Extraction de motifs spatio-temporels dans des séries d'images de télédétection : application à des données optiques et radar

Andreea Maria Julea

UNIVERSITÉ DE GRENOBLE

UNIVERSITÉ POLITEHNICA BUCAREST

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **STIC Informatique**

Arrêté ministériel : 7 août 2006

Et pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ POLITEHNICA BUCAREST

Spécialité: Ingénierie Électronique et Télécommunications

Présentée par

Andreea – Maria JULEA

Thèse dirigée par **Philippe Bolon et Vasile Lăzărescu**
codirigée par **Nicolas Méger**

préparée au sein du **Laboratoire LISTIC Annecy**
dans **l'École Doctorale SISEO**

Extraction de motifs spatio-temporels dans des séries d'images de télédétection - Application à des données optiques et radar

Thèse soutenue publiquement le **20 septembre 2011**
devant le jury composé de :

M. Teodor PETRESCU
Professeur UPB, Roumanie, Président
M. Jean - François BOULICAUT
Professeur LIRIS, INSA Lyon, France, Rapporteur
M. Alexandru BADEA
Directeur de Recherche, Agence Spatiale Roumaine, Rapporteur
M. Yannick BERTHOUMIEU
Professeur, IPB / ENSEIRB-MATMECA, France, Membre
M. Mihai DATCU
Professeur, DLR / UPB, Roumanie, Membre
M. Philippe BOLON
Professeur, LISTIC- Polytech Annecy, France, Membre
M. Nicolas MÈGER
Maître de Conférence, LISTIC- Polytech Annecy, France, Membre
M. Vasile LĂZĂRESCU
Professeur, UPB, Roumanie, Membre

Thèse

Extraction de motifs spatio-temporels dans des séries d'images de télédétection - Application à des données optiques et radar

pour obtenir

le grade de docteur

UNIVERSITÉ DE SAVOIE

Spécialité : *STIC Informatique*

UNIVERSITATEA POLITEHNICA BUCUREŞTI

Spécialité : *Ingénierie Électronique et Télécommunications*

par

Andreea Maria JULEA

Jury

Teodor PETRESCU	Président
Alexandru BADEA	Rapporteur
Jean-François BOULICAUT	Rapporteur
Yannick BERTHOUMIEU	Examinateur
Mihai DATCU	Examinateur
Philippe BOLON	Directeur de thèse
Nicolas MÉGER	CoDirecteur de thèse
Vasile LĂZĂRESCU	Directeur de thèse

Cette thèse a été préparée au Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (Annecy) et au Laboratoire d'Ingénierie Spatiale (Bucarest)

The noblest pleasure is the joy of understanding
Leonardo da Vinci

Remerciements

Les travaux présentés dans cette thèse ont été effectués en cotutelle entre l'Université de Savoie, France (Laboratoire d'Informatique, Systèmes et Traitement de l'Information et de la Connaissance - LISTIC) et l'Université "Politehnica" de Bucarest, Roumanie (Faculté d'Electronique, Télécommunications et Technologie de l'Information).

Je tiens tout d'abord à exprimer ma gratitude envers mes directeurs de thèse qui ont su guider et encourager mes recherches. Je souhaite remercier Nicolas Méger pour son énergie et son optimisme stimulants et son soutien constant dans cette aventure enrichissante. Je remercie également Philippe Bolon pour la qualité de son encadrement et ses conseils utiles. Je voudrais aussi remercier Vasile Lăzărescu pour m'avoir aidé à structurer mes idées et à développer une bonne démarche scientifique.

Mes remerciements les plus sincères vont aux rapporteurs Jean-François Boulicaut et Alexandru Badea pour leur lecture attentive de mon manuscrit de thèse, leurs remarques judicieuses et leur amabilité. Je remercie également les membres du jury, Mihai Datcu et Yannick Berthoumieu, de m'avoir fait l'honneur de participer à ma soutenance ainsi que pour leurs jugements très pertinents et constructifs. J'exprime ma plus grande reconnaissance à Teodor Petrescu, le professeur de mon premier cours concernant les satellites, pour l'honneur qu'il m'a fait d'accepter de présider mon jury de thèse.

Je souhaite remercier très chaleureusement tous ceux qui ont contribué au bon déroulement de ma thèse : le personnel de LISTIC et Politehnica et mes collègues de l'Institut de Sciences Spatiales.

Je tiens à remercier particulièrement Emmanuel Trouvé qui m'a fait découvrir le captivant domaine de la télédétection, pour ses idées et ses remarques originales.

Je tiens à exprimer toute ma reconnaissance à Christophe Rigotti pour son aide précieuse dans les problèmes informatiques et pour le temps consacré à nos dernières publications.

J'adresse mes vifs remerciements à Vasile Buzuloiu et Mihai Ciuc pour avoir guidé mes premiers pas dans le traitement d'images et dans la collaboration franco-roumaine.

Je remercie de tout cœur Inge Gavăt pour la confiance qu'elle m'a accordée, pour le soutien moral constant et pour sa gentillesse.

Un merci tout particulier à mes collègues stagiaires et doctorants rencontrés en France, Abdellah, Afaf, Alain, Alexandra, Alina, Amory, Anamaria, Azadeh, 2 Bogdan, Camille, Ciprian, Fabien, Florentin, Fred, Gabriel, Greg, Ivan, Karim, Lavinia, Mădălina, Maite, Mihaela, Raluca, Renaud, Sébastien, Selma, Sylvie, Yajing, Yoann, pour leur amitié, leur sympathie et les moments vécus ensemble.

Merci mille fois à mes amis et tout ceux qui m'ont soutenu par leur présence ou leurs messages d'encouragement le jour de la soutenance et plus particulièrement à Ana, Andreea, Ciprian, Corina, Daniela, Gabi, Iulia, Roxana, ...

Finalement je remercie affectueusement ma famille, spécialement ma mère et mon père, pour leur amour, leur support et pour leurs encouragements d'aller jusqu'au bout de ce projet de thèse.

Résumé

Les Séries Temporelles d'Images Satellitaires (STIS), visant la même scène en évolution, sont très intéressantes parce qu'elles acquièrent conjointement des informations temporelles et spatiales. L'extraction de ces informations pour aider les experts dans l'interprétation des données satellitaires devient une nécessité impérieuse. Dans ce mémoire, nous exposons comment on peut adapter l'extraction de motifs séquentiels fréquents à ce contexte spatio-temporel dans le but d'identifier des ensembles de pixels connexes qui partagent la même évolution temporelle. La démarche originale est basée sur la conjonction de la contrainte de support avec différentes contraintes de connexité qui peuvent filtrer ou élaguer l'espace de recherche pour obtenir efficacement des motifs séquentiels fréquents groupés (MSFG) avec signification pour l'utilisateur. La méthode d'extraction proposée est non supervisée et basée sur le niveau pixel. Pour vérifier la généricité du concept de MSFG et la capacité de la méthode proposée d'offrir des résultats intéressants à partir des SITS, sont réalisées des expérimentations sur des données réelles optiques et radar.

Mots clés : télédétection, Séries Temporelles d'Images Satellitaires, contraintes de connexité, motifs séquentiels fréquents groupés, images satellitaires optiques et radar

Abstract

The Satellite Image Time Series (SITS), aiming the same scene in evolution, are of high interest as they capture both spatial and temporal information. The extraction of this information to help the experts interpreting the satellite data becomes a stringent necessity. In this work, we expound how to adapt frequent sequential patterns extraction to this spatiotemporal context in order to identify sets of connected pixels sharing a same temporal evolution. The original approach is based on the conjunction of support constraint with different constraints based on pixel connectivity that can filter or prune the search space in order to efficiently obtain Grouped Frequent Sequential (GFS) patterns that are meaningful to the end user. The proposed extraction method is unsupervised and pixel level based. To verify the generality of GFS-pattern concept and the proposed method capability to offer interesting results from SITS, real data experiments on optical and radar data are presented.

Keywords : remote sensing, data mining, Satellite Image Time Series, connectivity constraints, grouped frequent sequential patterns, optical and radar satellite images

Table des matières

Table des figures

Liste des tableaux

Introduction

Le sujet de ce mémoire se trouve à la confluence d'un domaine, la télédétection satellitaire, avec des riches applications d'intérêt scientifique, économique et militaire et d'un domaine d'informatique en plein essor, l'Extraction de Connaissances à partir des Données (en anglais Knowledge Discovery in Database, KDD) (ECD). Les récents progrès de la technologie des capteurs satellitaires se manifestent par une croissance continue de la résolution spatiale, temporelle et radiométrique des acquisitions. Ceci engendre une augmentation du volume de données stockées telle qu'il est aujourd'hui nécessaire de faire appel à un traitement automatique permettant d'en extraire des informations utiles. L'exploration de données de ce type requiert l'utilisation de techniques permettant la découverte de relations, de modèles spatiaux, temporels ou spatio-temporels qui ne sont pas explicitement stockés dans les données.

Le traitement des données d'observation de la Terre permet l'obtention des motifs de la couverture terrestre et l'étude de la taille et de la dynamique de ces motifs. Le motif a la signification informatique d'une «expression dans un langage décrivant un sous-ensemble de données ou un modèle applicable à ce sous-ensemble» [70]. Les motifs révèlent une sorte d'«organisation» des variables dans le domaine spatial et/ou temporel résultant de la structure de la couverture terrestre et de son évolution [58].

Dans ce mémoire, nous proposons et présentons une méthode originale d'extraction de motifs (ou structures) spatio-temporels dans le contexte des Séries Temporelles d'Images Satellitaires (STIS). Une STIS est construite pour un même segment sol en agrégeant différentes acquisitions temporellement espacées. Les STIS confèrent une nouvelle dimension à l'observation de la Terre la dimension temporelle. L'analyse de telles données ne permet pas la simple réutilisation des outils dédiés au traitement d'images et requiert des techniques adaptées. Par ailleurs, l'exploitation des STIS constitue un enjeu majeur pour un nombre grandissant de domaines d'application liés à la compréhension de l'évolution de la couverture terrestre.

L'objectif de ce travail est l'introduction des concepts permettant la compréhension et la caractérisation des scènes dynamiques et, sur cette base, d'une méthode analysant conjointement les caractéristiques spatio-temporels des événements des STIS. La démarche se situe au niveau pixel pour préserver toute l'information à la haute résolution native. Ses caractères automatique et non-supervisé permettent d'envisager des applications de surveillance impliquant des flux importants de données et assurent une capacité d'adaptation à des situations nouvelles (catastrophes naturelles, renseignement militaire, etc.). Le but est de créer une représentation compacte des STIS décrivant le contenu informationnel de telle sorte à faciliter la reconnaissance de structures spatio-temporelles, c'est-à-dire la localisation spatiale et temporelle des phénomènes similaires.

Les images de télédétection peuvent fournir des informations spatiales comme les régions géographiques d'intérêt ainsi que des caractéristiques radiométriques. Lorsque ces images couvrent une même zone dans le temps, il devient possible de détecter des changements. La méthode traditionnelle d'analyse d'images de la Terre est une classification des pixels fondée sur

l'hypothèse que les pixels qui font partie de la même classe de couverture du sol sont proches dans l'espace des caractéristiques radiométriques [58]. L'analyse d'images de télédétection a beaucoup porté, durant les dernières décennies, sur l'analyse radiométrique d'images. Moins d'attention a été accordée à l'information spatiale capturée par ces images, alors que celles-ci constituent la base des efforts de cartographie et de modélisation dans les disciplines de l'environnement [59].

Dans les images satellitaires, plusieurs signaux sont enregistrés et associés aux coordonnées des pixels. Usuellement, les signaux sont les réponses des capteurs satellitaires à la radiométrie de la scène observée. Le pixel représente en fait une valeur moyenne dans chacune de ces trois dimensions : espace, caractéristique radiométrique et temps [198]. Le temps d'acquisition des capteurs satellitaires étant de l'ordre des microsecondes, sa moyenne a une influence négligeable pour la plupart des applications. Les moyennes sur l'espace et la caractéristique radiométrique sont importantes car elles déterminent la façon avec laquelle on peut observer les objets et repérer les différences radiométriques nécessaires pour identifier leurs propriétés.

L'information radiométrique, étant le signal utile, est systématiquement considérée tandis que le rôle joué par l'information spatiale ou temporelle dépend des méthodes de traitement. Dans le cas traditionnel de l'étude d'un nombre réduit d'images satellitaires, les coordonnées spatiales sont seulement utilisées pour la réalisation des cartes thématiques tandis que le temps constitue une valeur informative.

Pour une STIS, on doit tenir compte de l'échantillonnage temporel, la résolution temporelle étant importante pour la précision de description des évolutions des pixels. Les informations radiométriques et temporelles concourent à l'obtention des séquences d'évolution des pixels qui, après quantification, peuvent être considérées comme formant une base de séquences [10] représentant la STIS. Ces évolutions des valeurs des pixels dans une bande radiométrique réelle ou synthétique sont les éléments de caractérisation et de discrimination permettant de déceler les objets ou les phénomènes terrestres ainsi que leurs modifications dans l'espace et le temps. Nous proposons de caractériser ces objets et leurs évolutions à l'aide de motifs séquentiels [10].

L'extraction de motifs séquentiels est caractérisée par une quantité importante de données d'entrée, un espace de recherche exponentiel et un ensemble de solutions souvent trop grand [64]. Cette situation est préjudiciable parce que les performances se dégradent du point de vue de l'efficacité de l'exploration des données et de l'identification des fragments des connaissances intéressantes qui peuvent être estompés au sein de l'énorme quantité de motifs fournis.

Une autre problématique en techniques de fouille de données, en général, et dans la fouille de motifs en particulier, est la prise en compte des attentes des utilisateurs et de connaissance du domaine. Une solution à ces problèmes est offerte par le paradigme de l'extraction de motifs sous contraintes [64, 203, 111]. La fouille de données peut s'appuyer sur des contraintes qui représentent généralement l'intérêt de l'utilisateur et qui permettent de limiter les motifs trouvés à un sous-ensemble particulier satisfaisant ces contraintes. Une des contributions principales de cette thèse est la proposition d'une nouvelle contrainte et sa mise en œuvre au sein d'un processus d'extraction de motifs séquentiels. Cette contrainte, qui est une contrainte de connexité spatiale, peut en effet, en fonction de la mesure, être poussée complètement ou partiellement lors de l'extraction, et ceci en conjonction avec la contrainte standard de support (cette contrainte correspond dans notre cas à une contrainte de surface minimum). De cette façon, l'espace de recherche se trouve fortement réduit, les performances sont améliorées et moins de motifs sont proposés à l'interprétation de l'utilisateur. Ces motifs séquentiels satisfaisant la contrainte de support (surface minimum) et la nouvelle contrainte de connexité moyenne sont appelés motifs séquentiels fréquents groupés (MSFG).

Les principaux concepts considérés dans ce mémoire sont l'évolution temporelle comme

élément pour la caractérisation et la discrimination des pixels, et la connexité des pixels couverts par une séquence d'évolution. Ils permettent une prise en compte équilibrée des trois types d'information présents : radiométrique, temporel et spatial à partir de pixels voisins. La démarche présentée adopte une approche exploratoire sans hypothèse pré-formulée, capable de s'adapter à diverses applications (voir la partie III). La pertinence de la démarche proposée et la généricité du concept de MSFG sont vérifiées sur différents types de données, optiques et radar. Également, on vérifie la correspondance de motifs d'évolution identifiés (extraits selon leur support et connexité des pixels couverts) avec des entités réelles de la scène observée.

Organisation du mémoire

Ce travail est organisé en trois parties : I) l'état de l'art de la description spatio-temporelle de STIS, II) l'introduction des contraintes sur connexité et des MSFG, et le cadre théorique de leur extraction et III) des applications des MSFG sur des STIS différentes en nature et en résolution.

La première partie présente, de manière concentrique, l'ECD et les méthodes généralement disponibles, puis détaille certaines méthodes appliquées aux données satellitaires et, finalement, se concentre sur l'approche abordée et qui est appliquée aux STIS dans cette thèse : l'extraction de motifs séquentiels fréquents.

Le chapitre 1 offre une vision d'ensemble du processus ECD en présentant ses principales étapes et plus spécifiquement, le cas de la fouille de données avec prise en compte des caractéristiques spatiales et temporelles.

Le chapitre 2 présente des méthodes de traitement typiques des données de télédétection. De plus, sont présentées deux approches de fouille de données spatio-temporelles permettant d'extraire des évolutions de régions et des trajectoires d'objets.

Le chapitre 3 restreint la présentation à l'extraction de motifs séquentiels fréquents (MSF) dans des STIS : les fondements théoriques, les algorithmes d'extraction dédiés et l'introduction de contraintes dans le processus d'extraction. Sur la base de l'état de l'art sur l'ECD et l'analyse de STIS, les directions de la thèse sont explicitement formulées à la fin de la Partie I.

La deuxième partie constitue le noyau de ce mémoire avec l'introduction de nouvelles mesures de connexité, des contraintes associées et la mise en œuvre de ces différentes contraintes.

Dans le chapitre 4, sont détaillés les aspects théoriques des mesures de connexité locale, globale, moyenne et relative au support minimum. L'accent est mis sur la connexité moyenne dont la signification est accessible à l'utilisateur, qui peut être partiellement poussée (réduction de l'espace de recherche) et sur la base de laquelle est défini le concept de MSFG, et sur la connexité relative au support minimum, une mesure anti-monotone qui conduit à une extraction efficiente avec une forte réduction de l'espace de recherche.

Le chapitre 5 expose la mise en œuvre des techniques utilisant les contraintes de connexité : la simple vérification de la contrainte de connexité moyenne, l'implémentation au sein du processus d'extraction des contraintes anti-monotones de connexité globale et de connexité relative au support minimum, la relaxation de la contrainte de connexité moyenne par la contrainte anti-monotone de connexité relative au support minimum (CRSM) et la conjonction des contraintes de connexité moyenne et relative au support minimum.

Dans la partie III, les approches proposées sont vérifiées et évaluées par des expériences effectuées sur des données optiques et radar étudiées dans le cadre des projets ADAM [46] et EFIDIR [79]. Les résultats sont analysés de façon quantitative pour mettre en évidence l'influence des paramètres d'entrée sur l'extraction de motifs de différents types. De la même façon, une analyse qualitative de motifs résultés est réalisée au travers de leur interprétation et comparaison

avec des informations connues de scènes surveillées.

Le chapitre 6 décrit les expérimentations sur la STIS de la zone Fundulea, une région dont la plupart de la surface est couverte par de cultures agricoles. L'étude quantitative contient l'analyse des différents types d'extractions basées sur les contraintes de connexité proposées et l'étude qualitative est permise par l'existence d'une vérité terrain. L'étude atteste la qualité de MSFG extraits, assurée par un bon compromis entre les couvertures thématiques et les puretés de description.

Dans le chapitre 7 sont étudiées des données radar couvrant le lac Mead et la zone Chamonix Mont Blanc. Dans le premier cas, les motifs extraits de données interférométriques décrivent les déformations de la croûte terrestre suivant les variations du niveau de l'eau du lac et confirment leur capacité à trouver des modifications produites par des phénomènes non-aléatoires. L'extraction des motifs de données de polarimétrie radar, dans le deuxième cas, permettent la discrimination des propriétés intrinsèques des cibles terrestres. Les résultats préliminaires démontrent le potentiel du concept de MSFG. La partie III atteste ainsi de la capacité des MSFG à s'adapter aux spécificités des applications diverses.

Le bilan et les perspectives constituent la dernière partie du corps principal de la thèse. Ils offrent l'occasion de résumer les travaux et de mettre en évidence les avantages et inconvénients de l'approche proposée et de présenter des prochaines directions de recherche et des applications associées à la thématique de la thèse.

Finalement, quatre annexes présentent des sujets ponctuels apparus dans le développement de la thèse. Une annexe concerne des opérations de pré-traitement spécifiques aux données comme la réduction du nombre de symboles de valeurs de pixels et la réduction de dimensionnalité réalisée avec une transformation cosinus discrète. La deuxième annexe présente les méthodes de localisation spatiale et temporelle des motifs d'évolution extraits à partir de STIS, comme post-traitement général. La troisième annexe traite les résultats préliminaires de l'extraction de motifs séquentiels fréquents de longueur variable et complète, et des opérations de post-traitement spécifiques utilisées. La dernière annexe traite la phénoménologie de la scène STIS ADAM décrite avec l'Indice de Végétation Différentielle Normalisée (en anglais Normalized Difference Vegetation Index, NDVI) (IVDN), un indice approprié pour l'analyse des cultures agricoles, et un aspect d'altération d'information dû au sol nu.

Première partie

Description spatio-temporelle des Séries Temporelles d'Images Satellitaires : état de l'art

Introduction

Ces dernières décennies, de grandes quantités d'images couvrant de nombreux sites terrestres ont été acquises par les satellites, permettant ainsi la constitution de séries temporelles d'images satellitaires. De plus, les opportunités pour générer de nouvelles séries sont grandissantes : les satellites, de plus en plus nombreux, avec des résolutions spatiales et spectrales de plus en plus fines et des vitesses d'acquisition de plus en plus rapides, permettent l'augmentation de la fréquence de revisite d'une même scène. Ainsi, l'observation précise de la dynamique des scènes est de plus en plus accessible mais le volume disponible des structures spatio-temporelles devient énorme.

La nature et le volume de ces types de données dépassent les capacités humaines en termes d'analyse et d'interprétation. Par conséquent, l'intérêt d'appliquer des techniques d'extraction automatique de connaissances s'accroît. Pour répondre à cette problématique, le paradigme de l'ECD et de la fouille de données peut être appliqué aux STIS. L'extraction automatique des connaissances à partir des images satellitaires dans un contexte spatio-temporel est un défi majeur pour le domaine de la télédétection. Le processus de fouille de données dans une STIS implique la recherche de motifs spatio-temporels pertinents et utiles.

L'ECD est classiquement décrite comme un processus interactif et itératif de préparation des données, d'extraction de modèles/motifs à l'aide d'algorithmes de calcul, de visualisation et d'interprétation des résultats, lors d'interactions avec l'expert, afin d'obtenir de la connaissance. Les méthodes d'exploration proposent des solutions aux problèmes de recherche d'associations, de classification supervisée et non supervisée. Le chapitre 1 fait un passage en revue des étapes du processus ECD et présente un état de l'art des différents types de fouilles de données spatiales et temporelles.

Les STIS sont des données très riches pour l'étude de l'occupation des sols, soit pour la discrimination d'un état d'occupation des sols, soit pour l'étude de leurs évolutions ou des phénomènes de la surface terrestre. Pour l'exploration de données séquentielles offertes par les STIS, les méthodes de fouille de données doivent être ajustées d'une manière qui prend en considération la nature temporelle et spatiale des données. Le chapitre 2 expose un état de l'art de la problématique de l'analyse de STIS avec des moyens informatiques. Le niveau d'échelle des entités extraites (pixel ou objet), la nature supervisée - non supervisée de la démarche, les techniques de détection de changements, de clustering ou de classification et les formes de résultats de type motif local ou modèle global sont présentés et discutés pour faciliter le choix dans un cas réel. Le chapitre contient également deux exemples de fouille de données spatio-temporelles : structures spatio-temporelles et trajectoires.

Au niveau des motifs locaux, la fréquence des motifs est considérée comme le concept le plus utile pour mesurer le degré d'intérêt, assurant une certaine représentativité par le nombre minimal d'occurrences. Le chapitre 3 est une introduction au domaine de l'extraction de motifs séquentiels fréquents, MSF. Ainsi, on présente les définitions préliminaires, les travaux propres antérieurs du domaine et la nature combinatoire du problème. Sont exposés également les

différentes classes d'algorithmes spécialisés dans l'extraction de MSF. Ces méthodes d'extraction permettent d'identifier les ensembles de séquences ayant suivi la même évolution. De plus, elles permettent de caractériser cette évolution, en fournissant le motif partagé. Les principaux problèmes sont que le volume de motifs extraits est considérable et que la simple optimisation des algorithmes n'est pas suffisante pour réduire le nombre de motifs et les concentrer sur les attentes des utilisateurs [64, 202].

Le chapitre 3 présente les efforts faits pour solutionner ce problème. Les approches récentes utilisent des contraintes pour limiter le nombre et la portée des motifs découverts. L'utilisation de contraintes permet de concentrer le processus d'exploration dans des zones ou des sous-espaces où l'information utile est susceptible d'être acquise. Les contraintes permettent de coder les connaissances du domaine et l'intérêt de l'utilisateur dans le processus d'extraction. Ainsi, la contrainte constitue une dimension essentielle de l'extraction de motifs. Si les contraintes ont des propriétés de monotonie, elles peuvent être poussées en profondeur et le processus peut atteindre l'efficience et l'efficacité. Pour l'extraction de MSF c'est la contrainte anti-monotone de support (fréquence) qui est utilisée. L'extraction de MSF à partir de STIS a été introduite dans [123, 124, 114, 115, 125] sur des données optiques et radar en format mono et multi-canal. Le nombre de MSF extraits est assez grand et la dimension spatiale n'est pas exploitée. Pour réduire plus fortement l'espace de recherche il est recommandable d'utiliser des combinaisons avec d'autres contraintes anti-monotones. La partie finale du chapitre fournit l'occasion de discuter la problématique des contraintes et spécialement la gestion active ou passive des contraintes anti-monotones, seules ou en combinaison, dans un processus d'extraction de motifs séquentiels. On soutient qu'il est possible d'utiliser efficacement des algorithmes d'extraction de motifs séquentiels avec des contraintes appropriées, sur des données séquentielles, pour découvrir des informations pertinentes et intelligibles, en gardant le processus centrée sur l'utilisateur.

La partie II de cette thèse introduira une contrainte de connexité qui permettra l'implantation des caractéristiques spatiales dans le processus d'extraction. Cette contrainte, en fonction de la mesure choisie peut être partiellement ou complètement poussée au sein de ce processus.

Chapitre 1

L'Extraction de Connaissances à partir des Données - ECD

Sommaire

De nos jours, la société humaine est le témoin d'un développement sans précédent du volume et de la diversité des informations économiques, scientifiques et techniques. En revanche l'acquisition de ces informations, même concentrées par des outils informatiques dans des bases de données spécifiques, n'est pas suffisante ; elles doivent être converties en connaissances utiles. Ce chemin est devenu progressivement plus difficile en raison de l'explosion de données en vertu du développement technique. Le volume et la complexité de ces données exige l'aide de méthodes automatisées pour que des connaissances pertinentes puissent être obtenues.

Un domaine caractéristique de cette problématique est la télédétection satellitaire avec ses applications en surveillance environnementale, météorologique, climatique ou militaire. Le développement continu des techniques d'acquisition de données satellitaires (augmentation de la résolution, du nombre de canaux spectraux, de la fréquence de revisite, etc.) alimentent les bases de données avec une énorme quantité de données de divers types et attributs. L'automatisation du processus d'extraction de l'information devient une nécessité.

1.1 Le processus d'Extraction de Connaissances à partir des Données

Donner un sens à toute l'information contenue dans les données est illusoire voire inutile pour les chercheurs en informatique, mais également pour tous les utilisateurs. On suppose que ces données contiennent peut être des connaissances d'une grande valeur commerciale ou scientifique [111]. C'est en fait le postulat principal qui motive l'extraction de connaissances à partir des données. Une fois ce postulat admis, la question se pose de savoir comment des connaissances peuvent être extraites de ces données. L'opérateur humain ne peut pas traiter une telle quantité de données mais seul un expert humain peut évaluer les résultats d'une extraction. Le processus d'extraction de connaissances ne se limite donc pas à une extraction automatique. Il comporte plusieurs étapes pendant lesquelles l'expert humain doit faire des choix et évaluer les résultats en fonction de ses objectifs. Il peut passer à une étape suivante ou recommencer les étapes précédentes en utilisant une technique différente. De là découle la nature itérative et interactive de ce processus d'extraction.

L'ECD a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. Elle a été définie comme l'extraction d'une information implicite, non triviale, inconnue auparavant et potentiellement utile [77].

L'ECD est une discipline récente qui recoupe les domaines des bases de données, des statistiques, de l'intelligence artificielle et de l'interface homme/machine. Son objectif est de découvrir automatiquement des informations généralisables en connaissances nouvelles sous le contrôle des experts des données. Cela nécessite la conception et la mise au point de méthodes pour extraire les informations qui seront interprétées par les experts afin de les transformer, si possible, en connaissance.

Par rapport à ses domaines parents, l'ECD est caractérisée par le fait qu'elle extrait des connaissances pertinentes et intelligibles. Une connaissance pertinente a une valeur de vérité assez élevée ; on sait comment l'utiliser et elle s'accorde bien aux buts de l'utilisateur. Ainsi, la pertinence est presque complètement définie par l'utilisateur. Une connaissance est intelligible quand elle est exprimée dans le langage de l'utilisateur et avec la sémantique de celui-ci. Le fait que la connaissance découverte doive être auparavant inconnue limite en quelque sorte les buts et l'attente de l'utilisateur qui pourrait, par exemple, être heureux de retrouver quelque chose qu'il connaissait déjà (ce serait une forme spéciale d'intelligibilité) [133].

Un processus complet d'ECD met en jeu, de manière interactive et itérative, des multiples méthodes pour la préparation des données (le pré-traitement), leur exploration - la fouille de données, la visualisation et l'interprétation des résultats lors d'interactions avec l'expert (le post-traitement) [71]. Au coeur du processus se trouve l'étape de fouille de données qui consiste en l'application d'algorithmes d'analyse de données qui, sous les limites acceptables d'efficacité computationnelle, extraient, par exemple, les motifs locaux présents au sein des données. Les méthodes de fouille de données proposent des solutions aux problèmes de recherche des motifs locaux (règles d'association, motifs séquentiels), de classification supervisée et non supervisée. Les méthodes développées dans ce mémoire sont à base de motifs locaux. Compte tenu de la taille des bases de données, l'extraction de motifs locaux est un problème algorithmiquement ardu nécessitant la conception de méthodes efficaces pour parcourir l'espace de recherche.

Ainsi, l'objectif de l'ECD est de découvrir des motifs cachés, des tendances inattendues ou d'autres relations subtiles dans les données en utilisant une combinaison de techniques d'apprentissage automatique, des statistiques et des technologies de bases de données. Cette nouvelle discipline trouve aujourd'hui son application dans une gamme large et variée de scénarios d'affaires, scientifiques et techniques.

1.1.1 Données et pré-traitements

Le terme de données est utilisé pour désigner les faits constatés qui décrivent les états ou le comportement d'une entité, conformément à un ensemble d'attributs, denommés aussi champs ou variables, dont chacun correspond à une valeur particulière. Ces valeurs appartiennent à des ensembles spécifiques - les domaines d'attribut, qui représentent les valeurs qui peuvent être prises par l'attribut. En général, les domaines d'attribut peuvent appartenir à l'un des deux types : a) des valeurs réelles ou continues, sous-ensembles de nombres réels, où il y a une quantité mesurable dans une plage donnée et b) des valeurs catégorielles, ensembles finis de valeurs discrètes.

Il existe deux types d'attributs catégoriels : a) nominaux, où il n'y a pas d'ordre entre les valeurs, telles que les noms et les couleurs et b) ordinaux, indiquant qu'il existe un ordre parmi les valeurs, comme un attribut qui prend les valeurs basse, moyenne ou élevée. Lorsqu'il s'agit de transactions, deux types d'analyse peuvent être effectués :

- intra-transactionnelle, où l'analyse est effectuée entre les données traitées en même temps.
- inter-transactionnelle, où l'analyse est effectuée entre les données traitées à des instants différents.

L'analyse d'un comportement / évolution ne peut être effectuée par une analyse intra-transactionnelle, mais une analyse inter-transactionnelle est en mesure de le décrire.

En général, l'étape de pré-traitement est vue comme la préparation des données avant l'application de la fouille de données et le post-traitement comme l'évaluation et la présentation des informations découvertes à l'utilisateur final.

L'étape de pré-traitement consiste en un ensemble d'opérations effectuées sur les données afin d'améliorer leur qualité (par conséquent, les résultats de la fouille), et de réaliser leur mise en forme dans un format exploitable par les algorithmes de fouille de données. Le temps consacré à ce stade révèle la mauvaise qualité de la majorité des données existantes, et l'importance de ces opérations lorsqu'il s'agit de grands ensembles de données. Les opérations de pré-traitement peuvent être classées en quatre grands types de techniques : intégration de données, nettoyage des données, réduction des données [95] et transformation des données.

Les opérations d'*intégration de données* sont utilisées pour fusionner les données provenant

de plusieurs sources de données, potentiellement hétérogènes. Les principales difficultés sont liées au différents schémas de stockage et à l'existence des doublons.

Une fois que l'intégration des sources distinctes de données est atteinte, les opérations de *nettoyage* des données assurent la qualité des données. En général, trois situations distinctes sont traitées : les valeurs manquantes, les valeurs aberrantes ou le bruit et les incohérences dans les données.

En général, les bases de données contiennent de très grandes quantités de données, fait qui peut en découler du grand nombre d'enregistrements, du grand nombre d'attributs par enregistrement ou tout simplement de la complexité inhérente aux données. Étant donné que ces caractéristiques peuvent augmenter la difficulté du processus de fouille, la *réduction des données* est un besoin réel.

La réduction des données essaie d'obtenir une représentation réduite du jeu de données, plus petite en volume, mais qui produit les mêmes (ou presque) résultats analytiques.

La réduction des données comprend des techniques paramétriques et non paramétriques. Les techniques paramétriques supposent un modèle pour les données et tentent estimer les paramètres du modèle qui produisent un meilleur ajustement des données (par exemple, le modèle de régression), tandis que les techniques non paramétriques représentent, ou catégorisent, les données sans faire aucune hypothèse sur le modèle de données. Les principales méthodes non paramétriques utilisent les histogrammes, le clustering et l'échantillonnage des données.

Les principales stratégies pour la réduction des données sont la réduction de dimension, la réduction de numérosité, la discrétisation et la génération de hiérarchies de concepts.

L'analyse en composantes principales (ACP), les techniques de "multidimensional scaling" (MDS) [51], les cartes adaptatives de Kohonen (en anglais Self Organizing Maps) (SOM) [134] sont des outils classiques dans le contexte de la réduction dimensionelle. D'une manière générale, une fonction de coût (loss function) permet de construire les règles de projection de l'espace original des données vers l'espace cible de dimension plus faible. Pour les problèmes de classification, la conservation du voisinage apparaît comme un des aspects importants à maîtriser.

La discrétisation divise l'intervalle de valeurs possibles en sous intervalles. Elle est nécessaire dans le cas des algorithmes qui acceptent seulement des attributs catégoriels. Ainsi, en réduisant le nombre de valeurs d'un attribut, on fait la réduction du volume des données et la préparation pour de futures analyses.

La hiérarchie de concepts réduit les données en collectant et remplaçant les concepts de bas niveau (par exemple, l'amplitude) par des concepts de niveau d'abstraction plus élevé (amplitudes basses, moyennes ou élevées).

Autres techniques utiles dans le pré-traitement de données sont les transformations des données. Les plus générales méthodes sont le lissage, pour réduire le bruit, la construction de nouveaux attributs et la normalisation.

Quand il y a un grand nombre d'attributs, il est possible de sélectionner les plus pertinents. Cependant, parfois, les attributs existants ne sont pas en mesure de refléter la structure du domaine et la construction de nouveaux attributs peut aider à avoir un nouvel aperçu de la nature intime du problème (voir l'utilisation de l'IVDN pour la surveillance satellitaire des zones agricoles, chapitre 6). Cette construction est généralement obtenue par la combinaison d'attributs existants ou par la conjonction d'attributs booléens.

La normalisation est faite en échelonnant les valeurs possibles pour un attribut, de sorte qu'ils tombent dans un intervalle spécifié, habituellement de 0 à 1 (par exemple, l'IVDN). De

cette manière, des similitudes peuvent être détectées, en ignorant les différences d'échelle. Ce genre de transformation est appliqué à des valeurs continues, et il y a plusieurs stratégies pour réaliser la transformation.

D'autres approches utilisent des transformations comme la transformation Fourier discrète et la transformation en ondelettes pour compresser les données (voir l'annexe A).

Ainsi, une approche pour faire face à des séries temporelles est la traduction de la séquence originale dans une séquence composée de symboles nominaux. Il y a deux problèmes liés à cette traduction : choisir le domaine des nouveaux symboles - alphabet, et faire la traduction à partir des éléments à valeur réelle. Cette étape de pré-traitement est complexe et nécessite de faire de nombreux choix. De plus, il est difficile de déterminer a priori dans quelle mesure ces choix ont une influence sur le résultat des extractions. Une étude de l'influence de ce type de paramètre sur le rendement quantitatif et qualitatif de l'extraction de motifs séquentiels à partir de données réelles est réalisée dans ce mémoire (partie III, chapitre 6).

1.1.2 L'étape de fouille de données

La fouille de données est née du besoin d'exploitation de données produites, importées ou accumulées par un utilisateur, susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires. Plus précisément, la fouille de données concerne l'étape algorithmiquement difficile du processus d'ECD, qui produit des motifs locaux ou des modèles globaux potentiellement intéressants à partir des données préparées dans l'étape précédente.

Nous choisissons d'utiliser le mot «motif» avec la signification d'une condition sur un sous-ensemble des données, et utiliser le mot «modèle» pour la signification d'une condition sur tout l'ensemble des données.

Dans la phase de fouille de données, l'utilisateur doit choisir les modèles de représentation des données qu'il souhaite extraire (itemsets, règles d'association, clusters, etc.), définir les contraintes sur ces modèles et fixer les paramètres des algorithmes qui sont alors exécutés.

La fouille de données, héritière naturelle des statistiques, essaie d'aller plus loin, en fournissant en outre de modèles transformables en connaissances valides et exploitables, et des moyens automatiques pour classer et prédire les comportements futurs. Contrairement à la méthode statistique, la fouille de données ne nécessite pas que l'on établisse une hypothèse de départ qu'il s'agira de vérifier. C'est des données elles-mêmes que se dégageront les corrélations intéressantes, le logiciel n'étant là que pour les découvrir. La fouille de données adopte alors une démarche sans a priori (donc bien plus pragmatique) et essaie ainsi de faire émerger, à partir des données brutes, des inférences que l'expérimentateur peut ne pas soupçonner et dont il aura à valider la pertinence. La technique est particulièrement dynamique, car elle n'exige pas la préparation de requêtes.

D'après [93], les tâches générales de fouille de données peuvent être classées en deux catégories principales : descriptives (e.g. clustering, motifs locaux) et prédictives (e.g. classification non-supervisée, régression). La première identifie les motifs ou les relations dans les données, décrivant tout ou partie des données, alors que la dernière construit des modèles pour prédire le comportement de tout ou partie des futures/nouvelles données. Contrairement au modèle prédictif, le modèle descriptif sert à explorer les données, et non à prévoir de nouvelles données.

1.1.3 Le post-traitement

L'étape de post-traitement vise à accomplir deux tâches essentielles : analyser les résultats obtenus et présenter les meilleurs d'entre eux à l'utilisateur final. En substance, l'évaluation des motifs et modèles concerne trois aspects : la simplicité, la certitude et l'intérêt. Des motifs simples sont généralement préférés, car ils sont plus faciles à comprendre et parce qu'ils sont plus appropriés pour généraliser au-delà des cas connus. La certitude d'un modèle peut être indiquée comme mesure de la confiance que l'utilisateur doit mettre sur le motif. L'intérêt d'un motif s'évalue sur deux aspects : l'utilité et la nouveauté [17, 99, 98, 95, 94]. Un motif est utile s'il est facilement compris par les humains, valide sur de données nouvelles ou testées avec un certain degré de certitude et s'il répond aux besoins et exigences de l'utilisateur. Les mesures de nouveauté définissent la contribution de motifs à l'amélioration des connaissances sur le domaine. Les algorithmes d'extraction de motifs ou de construction de modèles permettent de découvrir des propriétés des données. Néanmoins, ces propriétés ne sont pas considérées comme de nouvelles connaissances tant qu'elles n'ont pas été interprétées et validées par un expert humain.

Les techniques de visualisation sont essentielles pour la présentation et l'interprétation efficace des résultats de l'exploration et même comme soutien pour le processus de fouille de données lui-même [130, 22].

La visualisation implique l'utilisation de techniques visuelles et graphiques pour représenter des informations, de données ou de connaissances. Ces techniques peuvent être employées dans les cas où des ensembles de données complexes doivent être expliqués ou analysés. L'idée essentielle est que les représentations visuelles peuvent aider l'utilisateur à obtenir une meilleure compréhension du contenu des ensembles de données, puisque le système visuel humain est plus enclin à traiter l'information visuelle que textuelle. Ainsi, les techniques de visualisation peuvent agir comme outils d'amplification des capacités perceptives, cognitives et analytiques des personnes pour leur permettre de résoudre des tâches complexes [13, 136, 112, 12].

1.2 État de l'art de la fouille de données spatiales et temporelles

On estime que 80 % des ensembles de données disponibles ont des composantes spatiales [69] et qu'elles sont souvent associées à des aspects temporels. Une telle quantité d'informations exige des techniques d'analyse adaptées.

La fouille de données spatio-temporelles est à la confluence de plusieurs domaines : les bases de données, l'apprentissage automatique, les statistiques, la visualisation géographique et la théorie de l'information. L'exploration de données de ce type est un nouveau domaine qui englobe les techniques pour découvrir des relations spatiales, temporelles ou spatio-temporelles utiles ou des modèles qui ne sont pas explicitement stockés dans des ensembles de données spatio-temporelles. Ces techniques s'occupent généralement des objets complexes avec des attributs spatiaux, temporels et autres. Les dimensions spatiales et temporelles ajoutent une complexité importante pour le processus d'extraction de données. Il y a une séparation traditionnellement appliquée à l'analyse des dimensions spatiales et temporelle : l'exploration des données temporelles (Fouille de Données Temporelles (en anglais Temporal Data Mining) (FDT)) [194], l'exploration de données spatiales (Fouille de Données Spatiales (en anglais Spatial Data Mining) (FDS)) [137, 162] et l'exploration de données spatio-temporelles (Fouille de Données Spatio-Temporelles (en anglais Spatio-Temporal Data Mining) (FDS-T)) [193].

Une application est généralement classée temporelle, spatiale ou spatio-temporelle selon le

problème cible à résoudre et la manière avec laquelle les ensembles de données sont collectées.

1.2.1 Fouille de données temporelles FDT

Un ensemble de données est dit séquentiel si ses données sont ordonnées à l'égard de certains indices, le temps étant l'exemple commun. Un ensemble de données séquentielles peut être fourni par toute grandeur physique «croissante», et sur cette base est décrite une «évolution» des états d'une entité. Ainsi, l'évolution d'un phénomène peut être décrite par la mesure des quantités appropriées pour l'augmentation monotone des valeurs d'une dimension sélectionnée, appelée «dimension d'ordre». Dans le cas de la STIS, le temps est la dimension fournissant l'ordre, et l'évolution est décrite au niveau d'une localisation élémentaire (pixel) par les grandeurs radiométriques mesurées dépendant des capteurs utilisés.

L'objectif global de la fouille de données temporelles est de découvrir les relations séquentielles ou des motifs qui sont implicitement présents dans les données et de fournir la possibilité d'explorer les aspects dynamiques des entités, au lieu de l'exploration de leurs caractéristiques statiques. En particulier, avec ce type d'analyse, il est possible de dégager certaines relations de cause à effet, ce qui permet la compréhension de l'évolution des entités analysées [194]. Les bases de données desquelles on peut extraire des motifs séquentiels sont de deux types. Elles peuvent être constituées d'une seule très longue séquence S [154, 156]. Dans ce cas, la fréquence d'une séquence S' peut être définie comme le nombre de fois où elle apparaît dans la séquence S. Une base de données peut être également constituée d'un ensemble de séquences, c'est le cas le plus courant de la "base de séquences" [6, 224, 178]. Dans ce cas, la fréquence d'une séquence S dans la base de données est définie uniquement comme le nombre de séquences de la base de données qui admettent S comme sous-séquence.

Les exemples de séries temporelles comprennent les données vocales, les historiques de prix des actions, les historiques de ventes, les enregistrements de tests d'un moteur, les données sismiques, les enregistrements de vols des avions, les données météo, les données environnementales, les données satellitaires, les données d'astrophysique, etc.

L'exploration de données temporelles a été fortement étudiée principalement pour les tâches (i) de prévision, (ii) de classification, (iii) de regroupement, (iv) de recherche et de récupération et (v) de découverte de motifs [95, 98]. Parmi les cinq catégories énumérées ci-dessus, les quatre premières ont été étudiées en détail dans l'analyse traditionnelle des séries temporelles et dans la reconnaissance des formes. Toutefois, les algorithmes pour la découverte de motifs dans de grandes bases de données sont d'origine plus récente et sont surtout discutés dans la littérature de fouille de données.

La tâche de *prédiction* (i) traite la prévision de futures valeurs de la série basée sur ses valeurs actuelles et passées. Habituellement, la prédiction exige la construction d'un modèle prédictif efficace pour les données (e.g. régression).

La *classification* (ii) suppose que certaines classes ou catégories aient été prédéfinies. L'objectif principal est d'identifier automatiquement pour chaque séquence d'entrée sa classe ou catégorie correspondante. Les techniques d'extraction de données temporelles pour la tâche de classification sont divisées en deux catégories [143] : méthodes fondées sur les modèles et méthodes basées sur les motifs. Les méthodes axées sur le motifs utilisent une base de données de séquences de caractéristiques comme prototypes des classes [68]. Pour n'importe quelle séquence d'entrée donnée, le classifieur fouille tous les prototypes en recherchant le plus proche ou semblable aux caractéristiques de la nouvelle l'entrée. Les méthodes fondées sur les modèles sont des techniques qui utilisent certains modèles puissants existants tels que les modèles de Markov

cachés, des réseaux neuronaux, machines vecteur support, etc.

Contrairement à la classification, le *clustering (regroupement)* (iii) ne prend pas en compte d'étiquettes de classe. Le clustering groupe l'ensemble des séquences d'après leur similitude. Le clustering est particulièrement intéressant, car il fournit un mécanisme dynamique pour trouver certaines structures (ou clusters) dans les grands ensembles de données.

Les techniques de *recherche et de récupération des séquences* (iv) jouent un rôle important dans l'exploration interactive de grandes bases de données séquentielles. Le problème consiste en la localisation efficace de séquences (souvent dénommées requêtes) dans les archives de grandes quantités de séquences (ou parfois dans une seule longue séquence). Dans certaines applications, il est possible d'estimer localement certaines caractéristiques symboliques (par exemple, les formes locales d'onde du signal) dans les séries temporelles à valeurs réelles et faire correspondre les séquences correspondantes symboliques [8]. Les approches de ce genre sont particulièrement pertinentes pour les applications de fouille de données car il y a beaucoup à gagner en termes d'efficience en réduisant les données de séries temporelles à valeurs réelles à des séquences symboliques, et en effectuant la mise en correspondance de séquences à ce nouveau niveau d'abstraction plus élevé.

Contrairement à des applications de recherche et de récupération, dans la *découverte de motifs* (v) il n'y a pas de requête spécifique disponible avec laquelle il faut interroger la base de données. L'objectif est simplement de découvrir tous les motifs d'intérêt. En ce sens, la découverte de motifs, avec son caractère exploratoire et non supervisé, est spécifique à la fouille de données.

La fouille de données temporelles est plus récente que l'analyse classique des séries temporelles, avec des contraintes et objectifs un peu différents. Une différence principale réside dans la taille et la nature des ensembles de données et dans la manière dont les données sont collectées [142]. Les méthodes de la fouille de données temporelles doivent être capables d'analyser des ensembles de données qui sont conséquents et les séquences peuvent avoir des valeurs nominales ou symboliques.

La deuxième grande différence (entre la fouille de données temporelles et l'analyse classique des séries temporelles) réside dans le type d'information qu'il est nécessaire d'extraire des données. Le domaine de la fouille de données temporelles s'étend au-delà de la prévision standard ou des applications de contrôle de l'analyse des séries temporelles. D'une plus grande pertinence peut être la découverte de motifs ou tendances utiles (et souvent inattendues) dans les données qui sont beaucoup plus facilement interprétables et utiles pour le propriétaire des données.

Dans toutes les applications de la fouille des données, la contrainte principale est le volume important de données. Il y a donc toujours un besoin pour des algorithmes efficaces. Améliorer la complexité en temps et espace des algorithmes est un problème qui continuera à attirer l'attention. Une autre question importante est celle de l'analyse de ces algorithmes afin que l'on puisse évaluer l'importance des motifs ou des règles extraits dans un certain sens statistique.

1.2.2 Fouille de données spatiales FDS

La FDS est aujourd'hui un domaine bien identifié de la fouille de données. Elle est née du besoin d'exploitation, dans un but décisionnel, de données à caractère spatial produites, importées ou accumulées et susceptibles de délivrer des informations ou des connaissances par le moyen d'outils exploratoires [227]. Sa principale caractéristique est qu'elle considère les relations spatiales de voisinage [66], car les attributs des voisins d'un objet d'intérêt peuvent avoir une influence significative sur l'objet lui-même. Le cadre général utilisé pour l'extraction de données

spatiales repose sur les relations de voisinage spatial entre les objets, sur des graphes de voisinage induits et sur des chemins de voisinage qui peuvent être définis par ces relations de voisinage [67]. Ces relations sont à l'origine implicites et nécessitent des jointures coûteuses sur des critères spatiaux pour être exhibées.

Les *données spatiales* sont généralement reliées aux objets caractérisés par une localisation spatiale et par plusieurs attributs non spatiaux. La *base de données spatiales* stocke des *objets spatiaux* représentés par des types de données spatiales et des *relations spatiales* entre ces objets [193]. Le défi crucial de la fouille de données spatiales est l'efficacité des algorithmes d'exploration en raison du volume de données et de la complexité des types de données et des méthodes d'accès.

Il est nécessaire de faire une distinction entre la fouille de données géographiques d'une part et le domaine étroitement lié de la fouille de données spatiales. Le terme «spatial» concerne les phénomènes ou les objets qui peuvent être incorporés dans un espace formel qui génère des relations implicites parmi des objets. «Géographique» désigne le cas particulier où les données objets sont géoréférencées et l'espace d'incorporation se rapporte (au moins conceptuellement) aux emplacements sur la surface terrestre.

Les données géographiques présentent souvent des propriétés de *dépendance spatiale* et d'*hétérogénéité spatiale*. La *dépendance spatiale* est la tendance des observations qui sont plus proches dans l'espace géographique à présenter des degrés supérieurs de similitude ou dissimilitude (selon les phénomènes). La proximité peut être définie en termes très généraux, tels que la distance, la direction et la topologie. L'*hétérogénéité spatiale* est souvent évidente puisque de nombreux processus géographiques sont locaux : les paramètres globaux ne reflètent pas bien le comportement d'un phénomène à un endroit particulier. L'hétérogénéité et la dépendance spatiale peuvent constituer un manque de spécification (comme les variables manquantes) mais peuvent aussi refléter la nature intrinsèque du processus géographique. De toute façon, ces relations sont porteuses d'information [163, 199].

Dans les applications typiques d'ECD, les objets sont discrets et peuvent être réduits à des points dans un espace multidimensionnel sans perte d'information. En revanche, les nombreuses entités spatiales ou géographiques, étant incorporées dans un espace continu, ne peuvent être réduites à des objets constitués de points sans perte d'information importante. Les caractéristiques telles que la taille et la morphologie des entités géographiques peuvent avoir des influences non négligeables sur les processus spatiaux ou géographiques.

Les hautes résolutions spatiales, temporelles et spectrales des systèmes de télédétection et des autres dispositifs de surveillance environnementale réunissent de grandes quantités d'images numériques géoréférencées.

Il est difficile pour les méthodes traditionnelles d'analyse des données, qui reposent principalement sur des opérations statistiques, de rendre compte de la diversité en types et en attributs de tels volumes de données.

La fouille de données spatiales peut être utilisée pour la compréhension des données spatiales, la découverte des relations entre les données spatiales et non spatiales, la construction de bases de connaissances spatiales, l'optimisation des requêtes, la réorganisation des données dans des bases de données spatiales, la saisie des caractéristiques générales de manière simple et concise. Ainsi, la technique peut être appliquée pour la détection, la cartographie et la prédiction de tout phénomène qui manifeste une composante spatiale.

Les techniques de la FDS comprennent la *classification* spatiale, l'*association* spatiale, le *clustering* spatial, l'*analyse spatiale des valeurs aberrantes* et la *prédiction* spatiale ([70, 162, 161]).

1.2.3 Fouille de données spatio-temporelles FDS-T

La disponibilité d'un très gros volume de données géospatiales, souvent continuellement mises à jour (par exemple des données de télédétection), met à l'épreuve la capacité de traiter les données et d'acquérir des connaissances utiles pour avoir la meilleure prévision et description de processus spatiaux ou temporels. La «spatialité» et la «temporalité» sont des caractéristiques essentielles pour la compréhension des processus de la surface terrestre. Plus récemment, l'intérêt de nombreux utilisateurs s'est porté sur la découverte des relations dynamiques et la compréhension des changements temporels. Les aspects spatiaux et temporels des données sont étudiés conjointement car ils sont souvent étroitement liés.

La dimension temporelle introduit une complexité supplémentaire dans le processus d'ECD. Une stratégie simple qui traite le temps comme une dimension spatiale supplémentaire n'est pas suffisante. Le temps a une sémantique différente de celle de l'espace : il est directionnel, a des propriétés uniques de mise à l'échelle et de granularité et peut être cyclique. Les modèles de données spatio-temporelles proposés (par exemple, [188, 220]) intègrent le temps et l'espace comme les principales dimensions. Cependant, les méthodes classiques de fouille de données ne reconnaissent pas le caractère unique des dimensions spatiales et temporelles. Les techniques d'extraction de données appliquées à des ensembles de données géographiques utilisent généralement des représentations très simples des objets géographiques et de la relation spatiale [34]. Les techniques de fouille de données devraient être modifiées pour exploiter les relations spatio-temporelles incorporées dans les ensembles de données.

En parallèle à la définition de Koperski [137] de la fouille de données spatiales, la fouille de données spatio-temporelles se réfère ici à l'extraction de connaissances implicites, des relations spatiales et temporelles, ou d'autres motifs qui ne sont pas explicitement stockés dans les bases de données [188, 220, 219, 107, 130]. C'est un sous-domaine de la fouille de données et de l'ECD, qui a débuté en informatique et en technologie de l'information dans les dernières décennies et qui pénètre maintenant dans presque tous les environnements de données complexes. Dans le secteur de la géoinformatique, l'étude de l'exploration de données spatio-temporelles a débuté récemment [193].

L'objectif général est d'étudier le comportement de certains objets (événements, entités, positions) dans l'espace ou dans le temps. Ainsi, les motifs résultants peuvent décrire des évolutions ou des trajectoires. Certains objets sont dynamiques, ils peuvent apparaître et disparaître ou changer la forme ou de taille. Cela complique la tâche de la technique d'extraction.

En s'appuyant sur les modèles qui ont été développés pour les données spatiales et temporelles, on peut classer les applications en quatre catégories [219] :

1. Des applications statiques où le temps ne fait pas partie des données enregistrées. Dans ce cas il est impossible d'extraire des modèles qui incluent la dimension temps. Il s'agit de toutes les applications d'exploration de données purement spatiales et, si nécessaire, le temps peut être retracé seulement par des informations externes de la construction de la base de données.
2. Des applications où les données sont enregistrées comme des séquences ordonnées d'événements selon des relations spécifiques comme avant et après, ou des relations plus complexes décrites comme rencontre, chevauchement, contemporain, etc.
3. Des applications où les données statiques sont enregistrées et horodatées à des intervalles de temps plus ou moins réguliers.
4. Des applications pleinement temporelles où la dimension temps est entièrement intégrée dans les données enregistrées (des événements, des processus, etc.).

Chapitre 2

Etat de l'art de l'analyse des STIS

Sommaire

L'ajout de la dimension temporelle à l'observation satellitaire de la Terre ouvre un grand nombre d'applications : analyse d'écosystèmes profondément affectés par l'activité humaine, évaluation de l'influence d'une guerre, de changements politiques, d'une sécheresse, d'un feu, d'une inondation, suivi de l'évolution de cultures et de l'occupation des sols. Ainsi, les STIS sont selon [187] une mine d'or en ce qu'elles sont des données à grande échelle temporelle et spatiale contrairement aux données terrain habituellement utilisées. En effet, l'observation régulière de la Terre permet un apprentissage des évolutions et changements normaux et permet donc, par complémentarité, de détecter les changements anormaux. Les développements techniques permettent de profiter de la croissance continue de la résolution spatiale et de la fréquence de revisite d'un même site. Ainsi, le nouveau type de données des STIS de haute résolution (STIS-HR) devient plus riche en information et très complexe, ce qui rend l'interprétation visuelle laborieuse et requiert des analyses automatiques.

Dans ce chapitre sont présentés quelques aspects caractéristiques de l'analyse de STIS souvent rencontrés dans les travaux des dernières années. Une série d'images satellitaires et les informations adjacentes offrent l'opportunité d'extraire des changements ponctuels ou des évolutions globales au niveau des entités de la scène étudiée et de caractériser les composantes de celle-ci. Différentes approches sont discutées en fonction du niveau de l'entité étudiée (pixel ou objet), de la nature supervisée ou non supervisée des démarches, des méthodes utilisées pour répondre aux diverses tâches (détection de changements, clustering ou classification) et de la nature et complexité de leurs résultats (motifs locaux, modèles globaux). Pour finir, deux types spécifiques de fouille de données sont analysés : la fouille d'information dans les images et la fouille de trajectoires.

2.1 Extraction des caractéristiques au niveau pixel et au niveau objet

Les facteurs qui influencent le choix entre l'analyse au niveau pixel ou au niveau objet sont :
- le rapport entre les dimensions de pixels et des entités d'intérêt de la scène ;
- le niveau de bruit des données ;
- l'objectif : classification thématique générale ou gestion détaillée des évolutions.

2.1.1 Extraction de motifs au niveau PIXEL

La méthode traditionnelle d'analyse des images de la terre est la classification (supervisée et non supervisée) des pixels fondée sur l'hypothèse que chaque pixel d'image est alloué à une seule classe (les pixels sont purs) et que les pixels qui capturent la même classe de couverture de la terre sont proches les uns aux autres dans l'espace des caractéristiques. L'hypothèse sous-jacente de cette approche est que les pixels d'une image se rapportent à des classes de couverture terrestre qui sont relativement séparables par leurs valeurs ou évolutions spectrales. Cette hypothèse n'est pas toujours valable, par exemple dans le cas où le pixel est trop grand par rapport à la variabilité des objets dans le paysage.

En raison de la non-correspondance de la grille de l'image avec les limites de l'objet réel, certains pixels mixtes (mixels) apparaissent dans l'image satellitaire. En cas de pixels mixtes, les réponses spectrales pures des différents objets de la scène sont confondues, menant à un problème de signatures composites. Les pixels mixtes ont été reconnus comme un problème affectant l'utilisation efficace des données de télédétection dans les méthodes de classification et détection des changements [52, 75, 41].

Fisher [72] a résumé quatre causes du problème de pixel mixte : (1) les limites entre deux ou plusieurs entités cartographiques, (2) la transition à l'intérieur des phénomènes cartographiables, (3) les objets linéaires subpixel (une route) et (4) les petits objets subpixel (une maison, un arbre).

L'***analyse au niveau pixel***, en préservant la résolution d'observation initiale et l'information originale de télédétection, assure une caractérisation plus détaillée des structures étudiées. Elle est une méthode indépendante du domaine d'application qui est très adéquate pour la gestion des évolutions de la couverture terrestre [72, 15]. On peut distinguer les modifications à l'intérieur des entités de la scène surveillée (intra-objets) et on peut offrir des informations détaillées en vue d'une aide à la décision. Par exemple, d'une scène agricole peuvent être extraites des informations agronomiques sur les cultures (rendement potentiel, risques de maladie, besoins en eau ou en azote, maturité) qui sont nécessaires pour la conduite optimale des cultures, pour estimer les productions et évaluer la qualité de l'environnement.

L'existence de pixels mixtes conduit à l'élaboration de plusieurs approches pour la classification 'soft' (souvent appelée floue dans la littérature de télédétection) dans laquelle chaque pixel est alloué à toutes les classes dans des proportions variables [197, 216, 76].

Les résultats de l'approche au niveau pixel sont sensibles à la résolution spatiale par l'intermédiaire de la proportion relative de pixels mixtes, ainsi qu'au bruit. Dans le cas d'une STIS, la méthode exige un recalage parfait des images.

Un exemple de traitement classique d'une STIS [87] présente une méthodologie et un ensemble de logiciels visant un regroupement pixel par pixel par une stratégie non supervisée et non hiérarchique. Chaque pixel est représenté par la série temporelle de ses valeurs ; les pixels caractérisés par des profils similaires sont affectés au même cluster selon un critère de distance minimale. Les clusters peuvent être regroupés ensuite conformément aux critères choisis par un analyste. Au lieu d'exprimer sa connaissance avant tout regroupement, l'analyste intéragit avec une partition exploratoire calculée automatiquement, ce qui offre un certain nombre d'éléments (répartition géographique, relation spatiale et les profils des classes de sortie) susceptibles de faciliter son jugement. Le résultat final est une image classifiée, où tous les pixels d'une même classe ont la même étiquette valeur d'octet (couleur).

Dans [168], la classification des STIS Radar à Synthèse d'Ouverture (en anglais Synthetic Aperture Radar, SAR) (RSO) au niveau du pixel est faite en utilisant des canaux synthétiques dans une étude sur l'estimation précoce des champs agricoles cultivés et non-cultivés. Les canaux synthétiques utilisés pour décrire les évolutions sont des fonctions mathématiques dans le temps comme les moyennes de valeurs de la rétrodiffusion ou les dates de rétrodiffusion maximale (canal qui contient des informations sur la préparation des champs agricoles et sur la phénologie de la récolte). Le même type d'application, toujours sur des STIS RSO, est présentée par [196] mais les auteurs utilisent aussi des canaux synthétiques plus raffinés. Le cycle phénologique des cultures est divisé en trois (au début, au cours et à la fin de la saison) et pour chaque période une moyenne de la rétrodiffusion est calculée. Le choix de telles fonctions est un a priori fort qui, pour certaines, efface l'idée même de transition, de changement d'état mais qui a l'avantage d'être synthétique. L'aspect spatial n'est pas pris en compte.

Dans [124, 123, 117, 125] l'évolution des valeurs des pixels est considérée comme essentielle pour caractériser le comportement de la couverture de la terre et les phénomènes météorologiques. L'approche [124, 123] introduit l'extraction des motifs fréquents d'évolution à partir de données des STIS en mono et multi-bandes optiques et radar. Les motifs séquentiels fréquents, MSF, sont extraits sous la contrainte de support, c'est-à-dire que le nombre de pixels couverts par un motif dépasse un seuil établi par l'utilisateur. Pour des détails, on peut consulter

l'annexe C.

Une approche d'extraction de motifs d'évolution de séries temporelles d'images satellites avec plusieurs bandes est présentée dans [185] et [186]. Reposant sur l'extraction de motifs séquentiels, la méthode a été spécifiquement conçue afin d'extraire des motifs d'évolutions non-majoritaires qui décrivent des changements. Les auteurs introduisent un seuil maximal de fréquence d'apparition, les motifs découverts ayant donc leur support compris dans un intervalle. Pendant le processus de découverte des motifs, ils éliminent les motifs contenant deux valeurs successives identiques sur une bande, étant intéressés surtout par les changements. Pour tous les exemples présentés ci-dessus, les relations de voisinage spatial des pixels ne sont pas utilisées, les pixels étant traités de façon indépendante.

2.1.2 Extraction des motifs au niveau OBJET

Une solution pour pallier les difficultés associées à la classification basée sur les pixels peut être de fonctionner à l'échelle spatiale des objets d'intérêt. Un objet est défini comme une entité caractérisée par un ensemble de paramètres dont les valeurs ne se modifient pas dans les différents points qui appartiennent à l'entité considérée. Plus simplement, on peut dire que l'objet a la propriété d'uniformité des paramètres de définition. Un des plus simples et des plus utilisés paramètres est la valeur du niveau de gris.

Par exemple, une approche basée sur l'objet diminue la possibilité de classer incorrectement les pixels individuels [100, 16, 189]. En effet, l'analyse au niveau objet est moins influencée par le bruit. Par la focalisation sur les objets du monde réel, les cartes produites de cette façon peuvent être plus directement utilisables par les analystes. L'analyse basée objets d'une séquence d'images exige la segmentation et la classification des images, le problème délicat étant la mise en correspondance des objets individuels dans le temps.

Dans [106] est présenté le cadre de la fouille de séquences d'images satellitaires météorologiques qui combine détection d'objets à partir des scènes et clustering des scènes. Les scènes sont d'abord classées automatiquement à l'aide de SOM en deux étapes. Les images incluant des objets similaires en mouvement sont affectées au même cluster. Puis les scènes contenant des objets proéminents sont aussi regroupées. Les groupes sont examinés et étiquetés sémantiquement par l'expert et les séquences d'images sont transformées en une base de données de séquences d'identificateurs de groupes. Ces séquences sont scrutées en utilisant des fenêtres glissantes sous des contraintes temporelles (maximum de temps écoulé entre les signatures) et de fréquence d'apparition pour déterminer des dépendances temporelles fortes de type épisodes comme $A \rightarrow B$. Ceci peut être lu comme "si on observe la signature A une ou plusieurs fois, alors, plus tard, on observe une ou plusieurs fois la signature B". En effet, il y a une condensation de la séquence (par exemple $A \rightarrow A \rightarrow B \rightarrow B = A \rightarrow B$) qui conserve seulement les changements, le rythme étant perdu [105]. Après l'extraction d'images qui incluent des objets proéminents basés sur le résultat du clustering, les positions et les formes des objets sont approximées en utilisant le modèle de mélange gaussien par l'algorithme Espérance-Maximisation (en anglais Expectation Maximization) (EM). Les objets identiques entre les scènes successives sont reconnus et étiquetés. D'autres connaissances telles que les trajectoires d'objets sont extraites à partir des séries temporelles d'identificateurs de groupes et des informations sur les objets. Finalement, les connaissances extraites sont stockées dans une base de données, qui permet des requêtes de haut niveau via l'interface utilisateur, et ainsi la découverte de connaissances par les experts du domaine est soutenue.

Dans [145], est présentée une méthode de segmentation spatio-temporelle d'une STIS à haute résolution qui consiste dans un premier temps en un partitionnement spatial de chaque image

et dans un second temps en une sélection temporelle d'instants pertinents. La scène de la STIS est considérée comme constituée de plusieurs couches, et l'intérêt est concentré sur l'arrière-plan dont les objets sont immobiles mais évoluent radiométriquement. On propose une représentation de la dynamique de cet arrière-plan dans un graphe d'adjacence temporelle des objets spatiaux (en anglais Spatial Object Temporal Adjancency Graph) (SOTAG). Les noeuds de ce graphe représentent les objets de la scène, et les arcs les correspondances temporelles entre ces objets traduisant une relation de type "devenir". Une navigation dans ce graphe permet donc d'accéder à l'histoire des objets. L'analyse de la STIS se décompose en trois étapes : la segmentation des images, la construction du graphe, et un regroupement des évolutions radiométriques similaires. La méthode de multi-segmentation jointe permet de ne pas découpler totalement la première et la deuxième étape. Une fois les objets extraits, les noeuds du graphe d'adjacence temporelle des objets sont déterminés, et les arcs peuvent alors être trouvés en estimant les correspondances entre objets. Une méthode permettant de regrouper de façon automatique les objets sous-jacents par similarité d'évolution radiométrique est proposée.

Dans [146] nous proposons une méthode qui permet d'exploiter à la fois les structures spatiales des STIS extraites par segmentation spatiale, et l'information temporelle des évolutions des pixels. Les cartes d'évolutions de la STIS, basées sur l'analyse au niveau du pixel, peuvent raffiner des segmentations réalisées sur chaque image de la série. Ainsi, les segmentations peuvent être enrichies par des informations concernant les évolutions temporelles (voir l'annexe C.5).

2.2 Méthodes usuelles d'analyse des STIS

2.2.1 Démarche supervisée et non supervisée

La raison d'être des méthodes supervisées est d'expliquer et/ou de prévoir un ou plusieurs phénomènes observables et effectivement mesurés. Concrètement, elles vont s'intéresser à une ou plusieurs variables de la base de données définies comme étant les cibles de l'analyse. Parmi les techniques développées dans ce but, on peut citer les techniques à base d'arbres de décision, les techniques statistiques de régressions linéaires ou non linéaires, les techniques à base de réseau de neurones (perceptron mono et multi couches, etc.), d'algorithmes génétiques, d'inférence bayésienne, etc.

Les méthodes non-supervisées permettent de travailler sur un ensemble de données dans lequel aucune des données ou des variables n'a une importance spéciale par rapport aux autres. Cela signifie qu'aucune variable n'est considérée individuellement comme la cible ou l'objectif de l'analyse. Par exemple, ces méthodes sont utilisées pour dégager un ensemble des groupes homogènes du point de vue des leurs caractéristiques. Dans ce but on peut utiliser des techniques à base de réseau de neurones (SOM), des techniques statistiques (classification ascendante hiérarchique, k-moyennes, le plus proche voisin, etc.), des techniques de recherche d'associations, etc.

Dans le cas particulier d'une STIS, après l'obtention des motifs d'évolution temporelle des pixels des images, les pixels montrant des motifs suffisamment semblables doivent être affectés à la même classe. Cela peut être fait :

- de manière *supervisée*, où les classes doivent être définies par l'analyste a priori. Un ensemble d'apprentissage doit être formé, constituant un ensemble de pixels pour lequel le motif et la classe d'attribution sont spécifiés.
- de manière *non supervisée*, où aucune classe n'est définie a priori. Les groupes sont générés par la technique elle-même. Les motifs sont regroupés selon leur similitude globale en groupes, dont le nombre peut être défini a priori par l'utilisateur ou déterminé par l'algo-

rithme lui-même. La similitude entre motifs peut être définie de plusieurs façons. L'objectif est de construire une partition (dans laquelle chaque motif est assigné généralement à une classe) telle que les classes sont en interne aussi homogènes que possible. Leur signification est dérivée par une interprétation appropriée des résultats.

La stratégie non supervisée est très présente dans la littérature. On peut citer Oja [170] : il est préférable de laisser le classificateur libre "étant donné que les catégories ne sont pas connues à l'époque où les extracteurs de fonctionnalités sont appliqués. Ainsi, la connaissance de la classe du modèle d'entrée n'est pas appropriée... ". Ou aussi Pao [172], qui insiste sur le fait que le classifieur doit "essayer d'identifier plusieurs prototypes ou exemples qui peuvent servir de centres de cluster". Un prototype peut être un motif réel ou un prototype synthétisé situé dans le cluster respectif.

2.2.2 Classification

La classification consiste à examiner les caractéristiques d'une entité et lui attribuer une classe en supposant que certaines classes ou catégories ont été déjà prédéfinies. Dans le cas d'une STIS, l'objectif principal est d'identifier automatiquement pour chaque séquence d'entrée sa classe ou la catégorie correspondante.

L'article [82] présente un processus de classification collaborative multi-stratégie multi-étape de données complexes. L'aspect collaboratif multi-stratégie est basé sur un raffinement automatique et mutuel des résultats de plusieurs classifications. Les auteurs ont défini un concept de résolution des conflits pour représenter les dissensions de classification qui utilise un critère de similitude, basé sur le recouvrement des classes. Le résultat fourni est unique et la méthode peut intégrer différents types d'attributs (numériques, symboliques ou structurés). La méthode proposée est appliquée pour une classification au niveau pixel d'images de télédétection.

Dans [140] est proposée l'utilisation d'un modèle dynamique pour améliorer la classification de la couverture du sol sur une séquence d'images de télédétection. L'approche consiste à représenter une parcelle de terre comme un système dynamique et à modéliser son évolution (en introduisant des connaissances sur les cycles des cultures) en utilisant le formalisme des automates temporisés. Afin d'affiner les résultats obtenus par un classificateur traditionnel, les observations données par une classification préliminaire des images sont combinées avec les états attendus fournis par une simulation avec un automate. Le document présente la modélisation capturée par le formalisme des automates temporisés et la méthode générale, qui repose sur des mécanismes de prévision et filtrage, qui ont été adoptés pour améliorer la classification d'une séquence d'images.

2.2.3 Clustering

Contrairement à la classification, le clustering est utilisé lorsqu'il n'existe pas de données étiquetées, ce qui signifie que c'est une opération sans supervision.

Dans le cas du clustering (regroupement), l'objectif est de regrouper les différentes entités dans des classes naturelles, des groupes (ou clusters) de sorte que les objets d'un même groupe soient aussi homogènes que possible et que deux groupes différents contiennent des objets suffisamment différents. Étant donné que chaque instance doit être semblable aux autres instances du même groupe, et dissemblable avec les instances des autres groupes, il est habituel que les méthodes de clustering fassent usage d'une mesure de similarité, afin d'identifier les groupes. Cette mesure de similarité, appelée aussi une fonction de distance, est indispensable pour effectuer le regroupement, mais peut être assez difficile à définir, en particulier en présence de types

de données complexes.

Une des principales difficultés est de découvrir le nombre de clusters. Après cela, il est nécessaire d'identifier ces clusters, d'attribuer un nouveau label pour chacun d'eux et de découvrir leurs descriptions lorsque cela est nécessaire. De cette manière, le regroupement est en mesure d'identifier des régions ayant des caractéristiques différentes, ce qui contribue à définir la répartition globale des données. Des corrélations entre les attributs peuvent également être trouvées, ce qui peut aider à la tâche de pré-traitement de sélection de caractéristiques.

Dans [131], les auteurs traitent le problème de segmentation des images satellites multi-dates en identifiant des clusters qui contiennent des pixels qui évoluent similairement dans le temps. Ils introduisent une mesure de similarité des séquences en tenant compte uniquement des changements d'états. L'information sur le rythme est perdue et l'information spatiale n'est pas prise en compte mais ils soutiennent que l'évolution au sol est saisie. L'algorithme utilisé repose sur la stratégie d'alignement des séquences discrètes (distance d'édition). L'intégration de la mesure de similarité dans le cadre d'algorithmes classiques de clustering est également discutée dans l'article.

Dans [81], une approche originale pour le clustering de données multi-dimensionnelles est proposée. La méthode est basée sur l'estimation du nombre de groupes à partir de la construction d'un arbre de représentation minimal (MST minimal spanning tree) avec l'algorithme Prim, afin de fournir les paramètres d'initialisation de l'algorithme K-moyennes classique. Les sommets sont supposés être répartis selon une distribution de Poisson et les mesures utilisées pour mesurer la similarité entre les points des données multi-dimensionnelles sont fondées sur des divergences informationnelles symétriques. Deux applications sont présentées en utilisant des mesures de réflectance à différentes longueurs d'onde. L'une porte sur la classification taxonomique des astéroïdes et l'autre concerne la segmentation dans une image multi-spectrale. L'aspect spatial n'est pas pris en compte.

2.2.4 Détection de changement

Dans une séquence d'images satellitaires, l'intérêt est focalisé sur la détection des changements entre deux images consécutives ou sur l'extraction des évolutions de la séquence entière. Les changements sont importants dans des applications telles que la surveillance de l'environnement et des forêts, le contrôle et la gestion de l'agriculture et l'extension des zones urbaines.

La détection de changements est le processus d'identification d'états distincts d'une zone en l'observant à des dates différentes. Ce processus est un type particulier de classification dédiée à la discrimination de deux classes de zones : "avec changements liés à un phénomène d'intérêt" et "autres". Ainsi, l'analyse est supervisée dans le sens que le type de changement doit être précisé et se limite en général à des données provenant de deux dates particulières. Le résultat final est une carte des zones de changement. Les techniques de détection de changement ont généralement besoin de renseignements sur le type de changement qui doit être pris en compte. Par exemple, on peut vouloir chercher des changements brusques, comme les inondations, les tremblements de terre, ou les catastrophes anthropiques (par exemple, [109]), ou on peut être intéressé par des changements progressifs tels que l'accumulation de la biomasse (par exemple, [214]).

La détection de changements peut se décomposer en deux étapes distinctes : l'obtention d'indices de changements suivie par la discrimination des zones de changements. Ces indices peuvent caractériser des changements ponctuels, dans le voisinage du pixel ou à l'echelle des structures.

Dans la première catégorie est faite la différence des attributs d'intérêt d'un même pixel entre deux dates [50], [151]. Bien qu'elles nécessitent un recalage fin et qu'elles soient sensibles au bruit, les techniques de détection de changement appliquées au niveau du pixel sont efficaces, en particulier lorsque des changements de réflectance sont évalués entre deux images optiques. Dans [28], les auteurs utilisent une série temporelle relative à la végétation, Indice de Végétation Amélioré (en anglais Enhanced Vegetation Index, EVI) (IVA), et détectent les changements en assignant à chaque pixel un score de changement. Ils proposent un algorithme récursif de fusion qui exploite les cycles annuels de végétation pour distinguer les points qui ont subi un changement des autres. La capacité de l'algorithme à ignorer les changements saisonniers naturels est particulièrement attractive. Parmi les limitations de l'algorithme, on peut mentionner qu'il n'utilise pas l'information spatiale qui est présente dans les données et qu'il ne découvre pas les motifs dominants dans les données. Une autre étude sur la détection de changements qui utilise des données Moderate Resolution Imaging Spectroradiometer (MODIS) de haute résolution est présentée dans [152]. La méthodologie de détection utilise des sommes annuelles d'un autre canal synthétique, l'IVDN pour un pixel du sol donné. Un changement est détecté si le z-score de la différence des sommes annuelles est supérieur à un seuil.

La deuxième catégorie prend en compte les changements des interactions spatiales des pixels. En général, des méthodes de caractérisation de texture sont employées. Par exemple, dans [149] est proposée une technique d'intégration des différences d'intensité et de texture entre deux images. La mesure de différence de texture repose sur la relation entre les vecteurs gradients. Elle est précise et robuste aux variations de bruit et d'illumination.

Concernant la troisième catégorie, les changements sont détectés sur les pixels qui changent de classe d'une image à l'autre. Par cette approche, on peut s'affranchir des problèmes d'étalonnage et de recalage des images mais reste le problème de l'extraction des structures d'intérêts (ce qui nécessite généralement des connaissances a priori). Une approche proposée dans [33], consiste à segmenter les deux images et fusionner les segmentations. Cette fusion consiste à obtenir une segmentation commune aux deux images, où chaque parcelle est homogène. Ensuite, chaque parcelle est caractérisée par un vecteur d'indices de changement qui permet d'évaluer les changements en restant à la résolution des parcelles. Cette méthode engendre des résultats détectant des zones de changements compactes et a pour avantage d'être robuste au bruit. Dans [27], les pixels sont regroupés en fonction de leur réflectance et de leur position pour trouver des objets. Les objets dont le comportement ne correspond pas à une référence stable sont sélectionnés.

On peut distinguer trois grandes familles méthodologiques pour l'analyse de changements [185]. Les méthodes bi-temporelles, permettent de situer et d'étudier des changements abrupts ayant lieu entre deux observations d'un phénomène à caractériser. Les méthodes correspondant à des techniques mixtes, principalement statistiques, s'appliquent généralement à deux images mais peuvent être combinées pour en analyser plusieurs. Les méthodes dédiées à l'étude de séries temporelles d'images sont généralement basées sur l'analyse de trajectoires radiométriques de pixels, afin de les comparer ou d'y détecter des ruptures.

Quel que soit le type de méthode de détection de changement utilisée pour l'analyse de STIS, il existe un décalage entre la quantité d'information que représentent ces séries temporelles, et la capacité des algorithmes à les analyser. Ces algorithmes sont le plus souvent dédiés à l'analyse bi-date d'une scène et se concentrent sur la cartographie des zones de changements et non sur leur caractérisation. Les méthodes bi-date sont de plus liées à des thématiques d'études spécifiques et sont incapables d'appréhender des changements ayant lieu au travers d'une STIS. Quant aux méthodes multi-dates, elles sont souvent difficilement interprétables et ne permettent pas de caractériser le changement.

Ces méthodes sont limitées par la dimension temporelle et ne permettent pas une extraction

d'information pour la gestion d'une base de séquences multitemporelles d'images. En outre, ces méthodes sont appropriées pour les changements abrupts mais sont peu performantes en cas de changements progressifs qui s'opèrent sur plusieurs images. Elles n'extraient pas l'information d'évolution disséminée au long d'une STIS. Pour la fouille de données, les intérêts sont plus vastes. Dans ce cas, on peut exploiter la totalité des données fournies par les images, sans une sélection a priori, et on peut obtenir une caractérisation totale des évolutions observées.

Les méthodes d'extraction de motifs séquentiels [10, 157, 225, 179] permettent d'identifier les ensembles de séquences ayant suivi la même évolution. De plus, elles permettent de caractériser cette évolution, en fournissant le motif partagé par cet ensemble de séquences. L'extraction de motifs d'évolution fréquents à partir des STIS, introduite dans [123, 124, 117, 125] utilise des données mono et multi-bande optiques et radar pour la météorologie et l'agriculture.

2.3 Représentation des données

Traditionnellement, la recherche en statistiques et apprentissage automatique a étudié des méthodes pour construire des modèles globaux, à savoir des synthèses descriptives de haut niveau de la structure générale des données en vue d'un certain objectif. Les exemples incluent des modèles statistiques de séries temporelles, des modèles de regroupement ou des modèles de classification comme les arbres de décision. La nature intrinsèque globale s'est avérée être le principal inconvénient que ces méthodes rencontrent dans des applications pratiques. Ayant un point de vue global sur les données, ces méthodes produisent rarement des perspectives nouvelles et surprenantes ; en effet, pour être valables, elles doivent résumer la plupart des données et, par conséquent, elles représentent des connaissances générales et évidentes pour les experts du domaine. Au contraire, ce qu'on recherche ce sont de connaissances intéressantes et surprenantes qui s'écartent du modèle de base déjà connu [24].

2.3.1 Motifs locaux

Les différentes techniques d'ECD peuvent être regroupées en deux catégories : l'extraction de motifs locaux et la construction de modèles globaux des données. Les premières visent à extraire des propriétés concernant des sous-ensembles des données alors que les techniques de construction de modèles sont globales et cherchent à mettre en évidence des propriétés de l'ensemble des données [53, 80].

Un motif est une structure physique ou abstraite d'objets. Il se distingue par un ensemble collectif d'attributs appelés caractéristiques [135]. Ainsi, un motif est une expression dans un langage décrivant un sous-ensemble de données ou un patron applicable à ce sous-ensemble [71].

L'extraction de motifs permet de répondre à des usages très divers. Les motifs obtenus peuvent soit être interprétés de manière brute (*motif local*), soit être combinés les uns avec les autres pour créer un *modèle global* (prédictif ou descriptif).

Les ***motifs locaux*** traduisent des situations précises au sein des données. Ils sont définis comme des régularités valables pour une partie des données. Le terme local désigne le fait qu'ils capturent certains aspects des données, sans donner une image complète de la base de données. Les motifs locaux ne représentent pas nécessairement les exceptions [97], mais des connaissances plutôt fragmentées et incomplètes, qui peuvent être assez générales (transmettant certains aspects des données). Ils offrent donc des informations qualitatives et locales enrichies éventuellement par la sémantique d'une *contrainte*, qui se révèlent facilement analysables de manière indépendante. Ces informations peuvent être complétées par une ou plusieurs mesures

statistiques.

Il y a un large éventail de méthodes pour découvrir les motifs d'intérêt potentiel pour l'utilisateur, mais les motifs les plus importants peuvent être perdus entre des informations trop triviales, bruitées et redondantes. Parmi les méthodes proposées pour réduire la collection de motifs, il y a les représentations condensées [40] ainsi que la compression de la base de données en exploitant le principe de Longueur de Description Minimale [129], celle du paradigme de la contrainte [169] ou l'approche de découverte d'ensembles de motifs [132].

À partir de la formalisation de [155] et la première application concrète à l'extraction des itemsets fréquents [29], plusieurs représentations condensées utiles ont été conçues. Les concepts de base utilisés dans les travaux récents sur la représentation condensés pour les itemsets fréquents sont : les ensembles fermés [226, 174], les ensembles δ-free [30, 31], les ensembles disjunction-free [35, 36], les ensembles généralisés disjunction-free [138], les itemsets non dérivables [38] et le cadre unifié [39, 40]. Les représentations condensées peuvent être étendues aux motifs séquentiels, par exemple les motifs maximaux [191] et les motifs fermés [218].

L'extraction sous contraintes centre la recherche d'informations suivant les souhaits de l'utilisateur. La contrainte la plus utilisée est la fréquence, le plus fondamental et en même temps le plus populaire type de découverte de motifs locaux étant la découverte (non supervisée) d'ensembles d'éléments fréquents [86]. Quand on parle de motifs fréquents, il est évident de penser à la fréquence comme à une mesure de localité : un motif très fréquent peut avoir un caractère global (c'est-à-dire il couvre une grande partie des données), un motif n'étant pas si fréquent est local (c'est-à-dire il ne couvre qu'une partie des données). Mais en même temps, il a besoin d'un certain support afin de se distinguer de la simple composante aléatoire. Du point de vue de la fréquence, la situation correspond à la définition de Hand dans la perspective classique de modélisation avec les motifs locaux [97] :

données = modèle de base + motifs locaux + composante aléatoire

Pour un niveau de support trop faible qui conduit à une puissance diminuée d'élagage de la contrainte de fréquence, l'espace de recherche peut exploser et le calcul peut devenir impossible. Cet effet peut être compensé par la puissance d'élagage d'autres contraintes que l'utilisateur peut exploiter pour restreindre la recherche de motifs intéressants. Une contrainte est non seulement utile pour élaguer l'espace de recherche, réduisant ainsi le calcul mais elle a également une valeur sémantique puisque le langage de contraintes est celui que les utilisateurs exploitent afin de définir quelles sont les tendances intéressantes.

L'importance des contraintes dans la recherche de motifs locaux est confirmée également par d'autres définitions. Selon A. Siebes "les motifs locaux sont décrits par des exigences structurelles, des attributs virtuels, et des conditions sur les valeurs d'attribut". De manière similaire, Boulicaut [165] affirme qu'un motif local est "une phrase d'un langage de motifs qui est a priori intéressante car elle répond à un ensemble donné de contraintes et raconte quelque chose sur une partie des données". Dans le cas des données de STIS, du fait de l'hétérogénéité des motifs spatio-temporels observés, la modélisation doit davantage s'orienter vers une description spatialement et temporellement localisée, plutôt que vers une description globale de la scène comme c'est le cas dans certaines recherches par le contenu d'images ou de vidéos.

Crémilleux et Soulet [54] proposent l'idée de contraintes globales pour écrire des requêtes traitant des motifs globaux comme un ensemble de motifs locaux. L'utilité des contraintes globales est de prendre en compte les relations entre les motifs locaux, ces relations exprimant une préférence d'utilisateur selon son attente. Ils proposent l'approche générique approximer-et-pousser pour l'exploration des motifs avec des contraintes globales.

2.3.2 Modèles globaux

Un ***modèle*** global est défini comme un sommaire des données à grande échelle, "une abstraction de la réalité". Il vise à décrire une caractéristique principale de l'ensemble de données.

Généralement, les modèles sont définis par un ensemble de paramètres estimés à partir des données. Souvent, il est possible de continuer à classer les modèles selon qu'ils sont prédictifs ou descriptifs. Les modèles prédictifs sont utilisés dans des applications de prédiction et de classification alors que les modèles descriptifs sont utiles pour le résumé des données. Par exemple, l'analyse d'autorégression peut être utilisée pour pédire les values futures d'une série temporelle en fonction de son passé. Les modèles de Markov constituent une autre classe populaire de modèles de prédiction qui a été largement utilisée dans les applications de classification de séquences [101], [89]. D'autre part, les spectrogrammes (obtenus grâce à l'analyse temps-fréquence des séries temporelles) et le clustering sont de bons exemples de techniques de modélisation descriptives. Elles sont utiles pour la visualisation de données et aident à résumer les données d'une manière commode.

Un champ actuel de recherche en plein essor est l'élaboration de modèles à partir de motifs locaux [165]. De manière générale, les modèles globaux sont issus d'un post-traitement sur les motifs locaux. Un défi consiste alors à rassembler les pièces du puzzle pour obtenir les ensembles de motifs satisfaisant une propriété impliquant plusieurs motifs locaux.

La construction de modèles issus de motifs locaux peut aussi tirer profit de la complétude des représentations provenant de l'extraction de motifs. Pour obtenir une véritable connaissance sur le domaine étudié, la construction de modèles nécessite l'utilisation de méthodes d'apprentissage automatique (classification, clustering) pour leur généralisation [24, 165].

Parce qu'un modèle global utile, comme un outil classificateur ou un modèle de régression, est souvent le résultat d'un processus de fouille de données, la question de comment activer les vastes collections de motifs en modèles globaux mérite attention. Un point commun à toutes les techniques de fouille de données pour obtenir des motifs locaux est qu'elles peuvent être considérées comme des techniques de construction des caractéristiques qui suivent des objectifs différents (ou contraintes). La redondance de ces schémas et la sélection de sous-ensembles convenables de motifs sont traitées dans des étapes distinctes, afin que chaque caractéristique qui en résulte soit très instructive dans le contexte du problème global de la fouille des données.

Knobbe [132] présente LeGo, un cadre générique qui utilise des techniques existantes d'extraction de motifs locaux pour une modélisation globale dans différentes tâches de fouille de données. LeGo commence par une phase d'extraction de motifs qui sont individuellement prometteurs. Les phases ultérieures établissent le contexte donné par la tâche globale de fouille de données en sélectionnant des groupes de motifs diversifiés et très informatifs, qui sont finalement combinés dans un ou plusieurs modèles globaux, des cibles globales de la fouille de données.

Dans [44], un modèle harmonique non linéaire est introduit pour identifier et prévoir la dynamique des classes de couverture du sol des écosystèmes naturels et anthropiques. Ce modèle à 5 paramètres s'ajuste remarquablement aux séries temporelles d'images satellite intra-annuelles de réflectance multispectrale et d'indices de végétation. Les attributs phénologiques peuvent être estimés avec précision à partir de la série temporelle ajustée et leurs dates et amplitudes peuvent être prédites par le modèle ajusté à seulement quelques observations antérieures.

2.4 Fouille d'information dans les images (Image Information Mining)

La fouille d'information est le processus engagé pour explorer et découvrir des connaissances à partir de grandes quantités d'informations stockées dans des bases de données. Dans cette approche, l'extraction d'information prend un caractère echelonnable, où l'information passe continument d'un niveau bas vers un niveau haut. Cette information est tout d'abord représentée par les primitives (couleurs, textures, formes), ensuite par la sémantique (forêt, ville, chien, visage, ...) et enfin par la connaissance (phénomène atmosphérique, visages souriants, ...). La sémantique étant liée à l'utilisateur, elle est prise en compte lors du processus de recherche en le faisant interagir avec le moteur de recherche. Aussi, le processus est généralement itératif et l'utilisateur signifie au moteur de recherche la pertinence de la recherche à chaque itération. En quelque sorte, par son intervention, le chercheur d'information fait apprendre au moteur de recherche une mesure de similarité adaptée à sa connaissance. Ce paradigme de la fouille de données présente l'avantage de combiner la flexibilité et la créativité de l'homme aux capacités énormes de stockage et de calcul des ordinateurs [89].

Dans le cadre de la gestion d'archives d'images satellitaires, Datcu et al. [57, 56] présentent le système Knowledge driven Image Information Mining (KIM) intégrant le paradigme de la fouille d'information. La méthode est basée sur la synergie de deux représentations de l'information, l'une objective et l'autre subjective. L'extraction des informations objectives est une approche guidée par les données, tandis que la partie subjective est centrée sur l'utilisateur. Dans un premier temps, l'information est extraite objectivement des données. Cette extraction d'information objective se fait en deux étapes majeures. La première étape consiste à extraire les primitives de chaque objet. La deuxième étape consiste à regrouper les primitives pour constituer des classes d'objets. Dans un second temps, la représentation subjective est obtenue à partir de la représentation objective par un apprentissage automatique sous les contraintes fournies par un utilisateur. Cette information est représentée par des modèles sémantiques et syntaxiques compréhensibles par l'utilisateur. Des techniques d'apprentissage supervisé sont mises en place pour guider l'algorithme vers les recherches les plus pertinentes. L'avantage d'un tel concept est qu'il est indépendant de la spécificité de l'application et s'adapte à la requête de l'utilisateur.

Les trois méthodes presentées ci-après suivent la structure mentionnée auparavant.

Le concept de modélisation de la trajectoire [102] est basé sur une modélisation bayésienne hiérarchique du contenu informationnel des STIS qui permet de lier l'intérêt d'un utilisateur à des structures spatio-temporelles spécifiques. La hiérarchie est composée de deux étapes d'inférence : une modélisation non supervisée des clusters dynamiques en générant un graphe de trajectoires qui code synthétiquement les structures spatio-temporelles, et une procédure d'apprentissage interactif basée sur les graphes, qui conduit à l'étiquetage (une classification) sémantique des structures spatio-temporelles. Des modèles stochastiques sont utilisés pour extraire des structures spatiales, spectrales et géométriques dans chaque image de la série temporelle au niveau pixel. Les graphes qui encodent les structures spatio-temporelles contenues dans la STIS sont ensuite inférés de ces structures. Enfin, en s'appuyant sur la représentation objective par des graphes, l'utilisateur final doit définir des exemples positifs et négatifs qui seront appris pour récupérer des structures similaires dans la STIS.

Dans [90] est traité le problème de l'extraction d'informations pertinentes à partir des STIS en s'appuyant sur le principe Information Bottleneck. La méthode repose sur la représentation objective de l'information par classification non supervisée et la sélection de modèles adaptés, couplée à une analyse débit-distorsion pour déterminer le nombre optimal de clusters. L'utilisation de cette méthode avec la famille de champs aléatoires paramétriques de Gibbs Markov est

présentée afin de découvrir et de caractériser les structures spatio-temporelles contenues dans des STIS.

Gueguen et al. [91] abordent le problème de la construction d'un index (ou dictionnaire) des bases de données objet comprimées en fonction du contenu en information. La méthode consiste à comprimer une base de données entière, l'index étant contenu dans le code. Les auteurs introduisent une mesure de similarité informationnelle basée sur la longueur des codes en deux parties et ils présentent une méthodologie pour la compression de la base de données en prenant en compte les redondances inter-objets et en utilisant cette mesure. L'index construit contient les informations minimales pertinentes suffisantes pour discriminer les données objets. Après, est présenté un codeur optimal en deux parties pour la compression des motifs spatio-temporels contenus dans des STIS. Ce codeur permet de mesurer la similarité, puis à calculer un index optimal d'événements spatio-temporels des STIS. L'index obtenu est représentatif de la teneur en information des STIS et permet des requêtes basées sur le contenu en information.

2.5 La fouille de trajectoires (Trajectory Data Mining)

L'omniprésence des technologies d'acquisition de localisation (dispositifs GPS, capteurs RFID, radars, les réseaux GSM, etc.) conduit à la constitution de grands ensembles de données spatio-temporelles et à l'opportunité de découvrir des connaissances utiles sur les comportements de déplacement des objets mobiles, qui favorise l'émergence de nouvelles applications et services.

L'analyse de ces données spatio-temporelles donne un aperçu du comportement des entités, en particulier, les schémas de migration des animaux. L'analyse des objets en mouvement a également comme domaines d'applications la géographie socio-économique, le suivi de véhicules, le sport (par exemple, les joueurs de football), l'analyse du trafic, l'analyse et le contrôle de la pêche, les prévisions météorologiques et l'analyse du mouvement (suivi des ouragans).

La trajectoire d'un objet en mouvement est typiquement une collection de signatures spatiales consécutives à des instants différents. La trajectoire est une collection d'arrêts d'un même objet se déplaçant à différentes localisations spatiales. La récupération des trajectoires similaires pourrait révéler des motifs sous-jacents de déplacement des objets dans les données. L'analyse des trajectoires peut être appliquée seulement si les trajectoires ont été fournies a priori.

Des motifs séquentiels fréquents représentant des sous-trajectoires des objets, c'est-à-dire des séquences de localisations, sont explorées dans [42]. Les trajectoires sont converties en lignes à plusieurs segments. Les segments similaires sont regroupés en utilisant une fonction de similarité qui tient compte de la proximité spatiale, basée sur l'angle et la longueur spatiale des segments. Finalement, les séquences fréquentes sont déterminées compte tenu d'un seuil de fréquence.

Dans [43], un seul objet est considéré et sa trajectoire est représentée comme une longue séquence d'événements à partir de laquelle des sous-trajectoires périodiques qui sont assez fréquentes sont extraites.

Dans [73], un motif est un groupe d'objets partageant un type de mouvement (direction, vitesse) à une date donnée dans une certaine région de l'espace. Cinq types de motifs de trajectoire basée sur le mouvement, la direction et la localisation sont proposés (convergence, rencontre, troupeau, leadership et récurrence). Dans [88] sont détectés les 4 premiers types de motifs définis dans [73] en utilisant des algorithmes de calcul approximatif. Les motifs spatio-temporels identifiés sont des sous-groupes d'objets ponctuels mobiles, avec des nombreux éléments localisés dans une région assez petite et présentant un mouvement similaire de point de vue de la direction, du but visé et/ou de la proximité.

Les auteurs présentent dans [167] une adaptation d'un algorithme de clustering basé sur la densité pour les trajectoires d'objets en mouvement, utilisant une notion de distance entre trajectoires. Ils mettent l'accent sur la dimension temporelle - essentiellement en élargissant l'espace de recherche des groupes intéressants en tenant compte des restrictions des trajectoires sources sur des sous-intervalles de temps. L'algorithme proposé de focalisation temporelle vise à chercher les intervalles de temps les plus significatifs, qui permettent d'isoler les groupes de qualité supérieure.

Dans [85], les auteurs proposent une extension du paradigme d'extraction de motifs séquentiels à l'analyse des trajectoires d'objets en mouvement. Ils introduisent les motifs de trajectoires comme des descriptions concises de comportements fréquents, en termes d'espace (les régions de l'espace visitées lors des déplacements) et de temps (la durée des déplacements). Ce travail est davantage axé sur des concepts de niveau supérieur (au lieu de découvrir un motif impliquant un endroit spatial précis, une localisation générale est trouvée). Ces localisations générales sont appelées régions d'intérêt (Regions-of-Interest ou RoI). Les motifs fréquents de déplacement entre ces régions sont découverts par la suite.

Toutes ces techniques pourraient être appliquées à des STIS pour analyser des trajectoires, après avoir identifié les objets et leurs déplacements spatiaux.

Les auteurs présentent dans [32] une approche automatique de haut-niveau pour la modélisation des connaissances spatio-temporelles à partir d'images satellitaires. Ils proposent d'utiliser une segmentation multi-approche comportant plusieurs méthodes de segmentation pour améliorer la modélisation et l'interprétation des images. Les expériences, sur deux scènes LANDSAT, montrent que leur approche surpasse les méthodes classiques de segmentation d'image et sont en mesure de prédire des changements spatio-temporels de couverture du sol.

Les deux articles suivants utilisent des données spatio-temporelles se composant de séquences d'événements localisés dans le temps et dans l'espace. Le contexte de motifs séquentiels peut donc être adapté plus facilement pour être appliqué dans ces cas.

Dans [107], les motifs séquentiels sont utilisés pour identifier des séquences significatives d'événements, où chaque événement est caractérisé spatialement et temporellement et appartient à un type spécifique d'événement. Par exemple, si un motif séquentiel "$A \rightarrow B$" est trouvé, alors il est interprété comme "des événements de type B ont tendance à se produire autour et après des événements de type A". Les auteurs utilisent un voisinage spatio-temporel décrit par une distance spatiale et un intervalle de temps fixés. Ils proposent, comme mesure d'importance pour les séquences spatio-temporelles, un indice calculé sur la base des densités des événements du voisinage spatio-temporel. Ils établissent l'interprétation statistique de cet indice à l'aide de statistiques spatiales.

Dans le contexte d'extraction de motifs fréquents, dans [210] sont trouvées des séquences fréquentes à partir des données spatio-temporelles. Par exemple, les lieux peuvent être des villes, et les données séquentielles peuvent être des enregistrements de température, humidité et pression. L'objectif est de trouver des motifs séquentiels fréquents dans les données. Un algorithme d'exploration en profondeur est proposé pour découvrir des motifs séquentiels à des localisations individuelles. Pour incorporer la dimension spatiale des données, l'algorithme examine les données à un niveau plus élevé de granularité spatiale en fusionnant certaines sous-régions voisines spatialement dans une région. Cette fusion est facilitée par le type de parcours spatial utilisé (la forme de la lettre Z parcourue en sens inverse pour des échelles croissantes).

Chapitre 3

Extraction de motifs séquentiels fréquents

Sommaire

La croissance rapide de la quantité de données numériques emmagasinées et les développements récents dans les techniques de la fouille de données ont mené à un intérêt croissant pour les méthodes d'exploration de données. L'Extraction des Structures Fréquentes est un de ces problèmes. Sa cible est la découverte de modèles structurés cachés dans les grandes bases de données. Les séquences sont la forme la plus simple de modèles structurés.

La recherche de motifs fréquents est un domaine important de la fouille de données et de la découverte de connaissance dans les bases de données. Son point de départ est lié à l'analyse de paniers d'articles et spécialement à la fouille de base de transactions dans le but de décrire le comportement des clients de supermarchés [6]. Un nombre important d'algorithmes a donc été proposé pour répondre à ce problème généralement connu sous le nom de fouille d'ensemble d'articles (itemset mining) dont les plus connus sont Apriori [9], ECLAT [221] et FP-Growth [96]. Ce problème a ensuite été étendu à la fouille de séquences [10, 205, 223, 178], permettant ainsi des applications dans la génomique ou pour l'extraction de motifs temporels, par exemple dans des réseaux de télécommunications ou dans la télédétection . Récemment, le problème a été étendu à des données encore plus complexes comme la fouille d'arbres fréquents et, plus généralement, de sous-graphes fréquents [64].

Les algorithmes d'extraction de motifs séquentiels visent à découvrir les séquences fréquentes existant dans une base de données. Les algorithmes sont pertinents quand les données à explorer ont une nature séquentielle, c'est-à-dire quand chaque morceau de données est une série ordonnée d'éléments, comme les événements dans le cas des informations temporelles. Le problème a été introduit initialement par [10]. L'objectif d'extraire des motifs séquentiels est de découvrir toutes les séquences fréquentes d'événements dans une collection de données.

Les motifs fréquents sont intéressants non seulement par eux-mêmes, mais ils sont également utiles pour d'autres analyses, y compris la classification et le clustering. Reflétant les fortes associations parmi plusieurs articles ou objets, les motifs fréquents capturent la sémantique sous-jacente dans les données. Ils ont été appliqués avec succès à des domaines interdisciplinaires au-delà de la fouille de données : la recherche d'indexation et de similarité des données structurées complexes, la fouille de données spatio-temporelles et multimédia, l'exploration des flux de données, la fouille du web et la fouille des erreurs de logiciels.

Ce chapitre présente la problématique d'extraction de MSF et les premières applications dans la télédétection. On introduit les définitions préliminaires, on présente la dimension d'extraction et les principaux algorithmes dédiés. On expose également les difficultés algorithmiques de l'extraction sous contraintes, les classes de contraintes qui en découlent et la nécessité d'exprimer des contraintes adéquates aux caractéristiques de la base de séquences de la STIS observée.

3.1 Motifs séquentiels dans les STIS

Les données séquentielles sont des données ordonnées [64]. La relation d'ordre établie sur ces données peut être temporelle, spatiale ou basée sur une autre grandeur physique unidimensionnelle. Par exemple, la longueur d'onde permet l'étude de la «signature» spectrale et une autre dimension spatiale, comme l'axe Z, permet l'investigation dans des données tomographiques).

On considère une séquence d'images couvrant une même zone géographique pendant une certaine période de temps. Cette séquence constitue une STIS. Chacune de ces images peut être vue comme un ensemble de pixels où la valeur d'un pixel indique la réponse radiométrique de la zone couverte par ce même pixel. Cette valeur dépend des longueurs d'onde auxquelles est sensible le capteur, qu'il soit passif ou actif, et des caractéristiques des objets terrestres. En considérant l'échelle des valeurs que peut prendre un pixel, il est possible de définir des intervalles

disjoints sur cette échelle, puis d'associer un symbole à chaque intervalle. Pour chaque image et pour chaque canal, à chaque valeur du pixel peut être associé un symbole en fonction de l'intervalle auquel elle appartient. De cette façon, l'impact des défauts de calibrage peut être réduit. Cette discrétisation et cet étiquetage permettent de ne pas forcément appréhender les données par leur valeur brute, et de les manipuler à des niveaux sémantiquement plus riches. Des informations supplémentaires concernant les méthodes de discrétisation utilisées dans ce mémoire pour les applications de fouille de données des STIS se trouvent dans l'annexe A.

On considère les valeurs d'un même pixel pour plusieurs images acquises à des dates différentes. À partir de ces données, il est possible de construire une séquence de plusieurs valeurs pour chaque pixel. L'ordre entre les différentes valeurs est donné par la dimension temporelle. Cette séquence de valeurs peut être traduite par une séquence de symboles selon le mécanisme de discrétisation et d'étiquetage évoqué ci-dessus. Au niveau de l'image, et en associant une telle séquence de symboles à chaque pixel, on obtient un ensemble constitué de milliers voire de millions de courtes séquences de symboles qu'il faut analyser afin de pouvoir extraire les motifs séquentiels. Il s'agit d'un contexte identifié en fouille de données, le contexte des bases de séquences.

Le contenu de ces deux derniers paragraphes constitue le point de départ pour la fouille de motifs fréquents d'évolutions à partir de STIS tel que nous l'introduisons dans [124, 123, 117, 125]. Nous proposons de faire usage des motifs séquentiels fréquents pour extraire automatiquement des évolutions, au niveau du pixel, qui sont contenues dans une série d'images satellitaires considérée comme une base de séquences. Les données satellitaires multi-temporelles utilisées, monocanal et multicanaux, proviennent de satellites météorologiques géostationnaires réalisés sous maîtrise d'oeuvre de l'Agence Spatiale Européenne (ESA) (METEOSAT) (bande optique et infrarouge thermique) et de la mission tandem radar European Remote Sensing satellite (ERS) (amplitude moyenne et cohérence interférométrique) pour une zone glaciaire du Mont Blanc [124, 123]. L'extraction de MSF de longueur variable est obtenue par application de l'algorithme Sequential PAttern Discovery using Equivalence classes (SPADE) [223] et des informations supplémentaires sont présentées dans l'annexe C. Dans [117, 125, 146], en utilisant un algorithme de type Trie on obtient l'extraction de MSF de longueur complète à partir des images Satellites Pour l'Observation de la Terre (SPOT) (3 canaux optiques) sur une zone agricole de Roumanie. Des détails sont présentés dans l'annexe C.

On introduit tout d'abord les définitions permettant de fixer le cadre général de l'extraction des motifs séquentiels dans une base de séquences [159] et nous poursuivons par l'analyse sur la dimension du problème.

3.1.1 Définitions préliminaires

Les bases de séquences sont des collections de séquences, qui elles-mêmes sont des successions d'événements. De façon plus formelle :

Définition 3.1. ***(Événements)*** Soit $E = \{i_1, i_2, ..., i_m\}$, un ensemble de m symboles distincts appelés *items* et muni d'un ordre total. Un *événement* (ou *itemset*) de taille l est un ensemble non vide constitué par l items provenant de l'ensemble d'items E, qui apparaissent ensemble. L'ensemble des événements est muni d'un ordre lexicographique défini à partir de l'ordre total des items.

Définition 3.2. ***(Séquence d'événements)*** Une séquence d'événements (ou séquence) de longueur L est une liste ordonnée composée de L événements $\alpha_1, ..., \alpha_L$ et représentée de la façon suivante : $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_L$.

Dans le contexte de séquences d'images, la séquence $C \rightarrow A \rightarrow A$ signifie qu'un pixel avait d'abord une valeur associée au symbole C, puis que sa valeur, dans l'image suivante, est passée à l'intervalle associé au symbole A pour rester à A dans la dernière image.

Définition 3.3. ***(Base de Séquences)*** Une base de séquences est un ensemble de couples (sid, seq) où seq représente une séquence et sid correspond à son identifiant (***s**equence **id**entifier*). De plus, pour toute séquence seq d'une base de séquences, chaque événement de seq possède un identifiant, noté eid (***e**vent **id**entifier*), qui peut correspondre à sa date d'apparition.

On note que la date eid peut représenter une position dans l'échelle d'une dimension physique, donc dans un ordre aussi. Pour une séquence d'images, cet eid peut représenter soit la date d'acquisition de l'image concernée, soit le numéro d'ordre de l'image dans la séquence d'images considérée. C'est cette dernière possibilité qui a été mise en pratique dans les expériences présentées dans la partie III. Une base de séquences peut être représentée sous la forme d'un ensemble de triplets $(sid, < eid, items >)$ avec $items$ la liste des items composant l'événement situé à la position eid dans la séquence sid. Ainsi, si on considère la base de séquences présentée ci-dessous, on peut construire sa représentation sous la forme du Tableau 3.1.

$((0,0), \langle(1,A),(2,B),(3,C),(4,B),(5,D)\rangle)$,

$((0,1), \langle(1,B),(2,A),(3,C),(4,B),(5,B)\rangle)$,

$((1,0), \langle(1,D),(2,B),(3,C),(4,B),(5,C)\rangle)$,

$((1,1), \langle(1,C),(2,A),(3,C),(4,B),(5,A)\rangle)$

sid	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,1	1,0	1,0	1,0	1,0	1,0	1,1	1,1	1,1	1,1	1,1
eid	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
item	A	B	C	B	D	B	A	C	B	B	D	B	C	B	C	C	A	C	B	A

TAB. 3.1 – Les valeurs sid, eid et $items$ de la base de séquences considérée

Cette base de séquence, dans le cas des images, correspond à une série de 5 images contenant 4 pixels. Par exemple, au niveau du pixel identifié $sid(0,1)$, les différents valeurs et/ou plages de valeurs successivement prises par ce pixel sont : B, A, C, B et B.

On définit maintenant les objets recherchés dans une base de séquences, c'est-à-dire les motifs.

Définition 3.4. ***(Motif)*** Un motif est une séquence extraite, et il est représenté de la façon suivante : $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$.

Les relations de généralisation et spécialisation sont définies par l'intermédiaire de sous- et sur-séquence (ou sous- et sur-motif)

Définition 3.5. ***(Sous-séquence)*** Une séquence (ou motif), de la forme $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ est appelée une sous-séquence (ou sous-motif) d'une séquence $\beta_1 \rightarrow \beta_2 \rightarrow ... \rightarrow \beta_m$ s'il existe des entiers $1 \leq i_1 < i_2 < ... < i_n \leq m$ tels que $\alpha_1 \subseteq \beta_{i_1}, \alpha_2 \subseteq \beta_{i_2}, ..., \alpha_n \subseteq \beta_{i_n}$.

Définition 3.6. ***(Sur-séquence)*** Toute séquence (ou motif) β ayant pour sous-séquence une séquence α est une sur-séquence (ou sur-motif) de α.

La relation $\alpha \Rightarrow \beta$ est une spécialisation et $\beta \Rightarrow \alpha$ est une généralisation. Par exemple, si l'on considère la séquence C $\rightarrow$ AD $\rightarrow$ A, alors le motif C $\rightarrow$ A $\rightarrow$ A en est une sous-séquence car $C \subseteq C$, $A \subseteq AD$ et $A \subseteq A$. De même, le motif $A \rightarrow A$ est une sous-séquence de $C \rightarrow AD \rightarrow A$ car $A \subseteq AD$ et $A \subseteq A$.

Définition 3.7. ***(Occurrence d'un motif, support)*** Si un motif α est sous-séquence d'une séquence γ d'une base, on dit que α apparaît dans γ. L'apparition d'un motif dans une séquence particulière d'une base est appelée occurrence d'un motif et correspond à une liste d'événements accompagnés de leurs identifiants dans cette séquence de la base. Une occurrence d'un motif $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ se représente de la façon suivante : $\alpha_1(eid(\alpha_1)) \rightarrow \alpha_2(eid(\alpha_2)) \rightarrow ... \rightarrow \alpha_n(eid(\alpha_n))$.

Le nombre d'occurrences d'un motif est alors défini comme le nombre de séquences dans lesquelles le motif apparaît au moins une fois. Ce nombre est appelé *support*.

Plus précisément le *support* d'un motif α dans une base de séquences BS est le nombre de couples (sid, seq), dans la base de données contenant α, c'est-à-dire,

$$supp_{BS}(\alpha) = |\{\langle sid, seq\rangle | (\langle sid, seq\rangle \in BS) \wedge (\alpha \subseteq seq)\}| \tag{3.1}$$

Définition 3.8. ***(Motif Séquentiel Fréquent, MSF)*** Soient BS une base de séquences et σ un entier positif appelé seuil de support absolu. Soit α un motif et $supp(\alpha)$ le nombre de séquences de BS dans lequel il apparaît. Le motif α vérifie la contrainte de fréquence minimum dans une base de séquences BS si α est une sous-séquence d'au moins σ séquences de BS, c'est-à-dire si $supp(\alpha) \geq \sigma$.

L'ensemble des techniques d'extraction de motifs s'appuie sur cette notion de *support* afin de sélectionner et d'extraire les motifs dits fréquents, c'est-à-dire les motifs dont le support est supérieur ou égal à une valeur notée σ. Cette notion de *support* pose de fait une contrainte sur les motifs nommée contrainte existentielle. L'utilisation active de cette contrainte anti-monotone est nécessaire aux différentes techniques d'extraction car elle permet de réduire l'espace des motifs envisagés durant le processus de calcul. Autrement dit, les motifs de base recherchés sont les MSF.

Le seuil de support peut être aussi spécifié comme un seuil de support relatif $\sigma_{rel} \in [0, 1]$. Alors un motif α est fréquent si le $supp(\alpha)/|BS| \geq \sigma_{rel}$, où $|BS|$ est le nombre de séquences complètes dans BS. Dans le jeu de données présenté auparavant, le motif séquentiel $A \rightarrow C \rightarrow B$ a les quatre événements suivants (les éléments dans un événement n'ont pas besoin d'être contigus dans le temps) :

$((0, 0), \langle(1, A), (3, C), (4, B)\rangle$,

$((0, 1), \langle(2, A), (3, C), (4, B)\rangle$,

$((1, 0), \langle(2, A), (3, C), (5, B)\rangle$,

$((1, 1), \langle(2, A), (3, C), (4, B)\rangle$

Le motif a quatre événements, mais apparaît dans seulement trois séquences d'évolution de pixel différentes. Ainsi son support est $supp(A \rightarrow C \rightarrow B) = 3$. Alors, si $\sigma_{rel} = 3/4$ *(seuil de support relatif)*, le motif est considéré comme étant un motif fréquent. Enfin, il faut remarquer qu'une étiquette peut être répétée dans un motif, et par exemple, le motif $C \rightarrow C$ a deux événements, l'un dans la troisième et l'un dans la quatrième séquence.

Afin de pouvoir manipuler de façon aisée les motifs et leurs propriétés lors de la présentation des différentes expériences, les définitions suivantes sont introduites :

Définition 3.9. ***(k-motif, taille, préfixe et suffixe d'un motif)*** Un motif $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ est un $k-motif$ si $\sum_{i=1}^{n} |\alpha_i| = k$ (il est composé de k items) et k est aussi appelé *taille* du motif. Le $suffixe$ d'un motif est le plus grand item (pour l'ordre total des items) contenu dans le dernier événement du motif. Le *préfixe* d'un motif est le motif privé de son suffixe.

Par exemple, le motif $A \rightarrow BC$ est un 3 - motif, i.e. de taille 3, ayant pour préfixe $A \rightarrow B$ et pour suffixe C.

Définition 3.10. ***(Longueur et largeur d'un motif)*** La longueur d'un motif $\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_n$ est n (le nombre d'événements qu'il contient) et sa largeur est $max_{i=1..n}|\alpha_i|$.

Si l'on reprend le motif $A \rightarrow BC$, sa longueur est de 2 et sa largeur est de 2.

Pour des définitions plus génériques et plus formelles on peut consulter [10].

Pour réduire l'espace de solutions de la fouille de motifs séquentiels fréquents, il est possible d'explorer et utiliser juste des motifs séquentiels fréquents maximaux ou fermés.

Définition 3.11. ***Motif séquentiel fréquent maximal*** Un motif séquentiel fréquent α, $(supp(\alpha) > \sigma)$, est maximal s'il n'y a aucun sur-motif approprié $\beta \supset \alpha$ tel que β soit fréquent (donc $supp(\beta) < \sigma$).

Définition 3.12. ***Motif séquentiel fréquent fermé*** Un motif séquentiel fréquent α, $(supp(\alpha) > \sigma)$, est fermé s'il n'y a aucun sur-motif $\beta \supset \alpha$ tel que $supp(\beta) = supp(\alpha)$.

L'ensemble de motifs séquentiels fréquents fermés est une compression sans perte du set de tous les MSF. Pour α un motifs séquentiels (MS), le fait si α est un MSF et l'information sur son support peuvent être dérivés de l'ensemble de MSF fermés comme suit :

- α n'est pas un MSF $(supp(\alpha) < \sigma)$ si et seulement s'il n'y a pas aucun MSF fermé β tel que $\alpha \subseteq \beta$
- si α est un MSF, alors $supp(\beta) = supp(\alpha)$ ou β est un MSF fermé tel que $\alpha \subseteq \beta$ et il n'y a aucun autre MSF fermé β' tel que $\alpha \subseteq \beta' \subset \beta$.

Si on considère une série temporelle d'images comme une base de séquences où chaque séquence trace l'évolution d'un pixel donné, il est possible de trouver toutes les évolutions fréquentes au niveau du pixel en extrayant tous les motifs séquentiels fréquents. On va voir, dans la section 3.1.2 que le nombre de motifs séquentiels est très grand. Par conséquent, la recherche de tous les motifs séquentiels peut être une tâche très consommatrice de ressources.

3.1.2 Analyse du problème

Soit une base de données avec $s = |E|$ le nombre d'items (symboles) différents possibles de l'ensemble E. Soit I l'ensemble des itemsets possibles. Son cardinal est :

$$|I| = \sum_{j=1}^{s} C_j^s = 2^s - 1 \tag{3.2}$$

où C_j^s donne le nombre d'itemsets possibles qui contiennent j items.

Pour comprendre le problème de l'extraction des motifs séquentiels, on commence en considérant que la base de données a des séquences avec au plus m itemsets et chaque itemset a au plus un item. Dans ces conditions, il y aurait s^m séquences différentes possibles avec m itemsets et

$$\sum_{k=1}^{m} s^k = \frac{s^{m+1} - s}{s - 1} \tag{3.3}$$

séquences différents de longueur arbitraire. De même, si chaque itemset a un nombre arbitraire d'items, il existe S_m séquences fréquentes possibles avec m itemsets, avec la valeur de S_m donnée par l'équation 3.4.

$$S_m = |I|^m = (2^s - 1)^m \tag{3.4}$$

En général, le nombre de séquences possibles est :

$$S = \sum_{k=1}^{m} S_k = \sum_{k=1}^{m} (2^s - 1)^k = \frac{(2^s - 1)^{m+1} - 2^s + 1}{2^s - 2} = \Theta(2^{ms}) \tag{3.5}$$

Dans le cas d'une base de données satellitaires avec b bandes spectrales, un événement peut contenir au plus b items, un item au plus pour chaque bande. Le nombre d'itemsets devient :

$$|I| = [\prod_{j=1}^{b} (s_j + 1)] - 1 \tag{3.6}$$

où s_j est le nombre de symboles utilisés pour décrire les valeurs de la bande j. Parce qu'il y a des motifs qui peuvent contenir un nombre de symboles plus petit que b, il faut ajouter une unité à s_j pour le cas où l'itemset ne contient pas de valeur pour cette bande. Ainsi, le nombre de séquences possibles devient :

$$S = \sum_{k=1}^{m} [(\prod_{j=1}^{b} (s_j + 1)) - 1]^k \tag{3.7}$$

Pour le cas d'une seule bande avec s symboles, la relation 3.7 se réduit à la relation 3.3 :

Même pour une seule bande, avec $s = 3$ symboles et un nombre d'images $m = 20$, le nombre d'évolutions possibles dépasse 5 milliards.

La vérification de l'ensemble de ces motifs n'est évidemment pas traitable lorsqu'on envisage des données réelles. En outre, si aucun critère supplémentaire n'est utilisé, les utilisateurs finaux auront à interpréter trop de motifs séquentiels. La fréquence des motifs est considérée comme le premier concept utile pour mesurer le degré d'intérêt, assurant une certaine représentativité par le nombre minimal d'occurrences. Par conséquent, dans une première étape il est proposé de sélectionner les motifs fréquents.

3.2 Algorithmes d'extraction de motifs séquentiels fréquents

De nombreux algorithmes ont été conçus pour effectuer les tâches d'extraction de motifs séquentiels fréquents. Les trois approches principales sont : les approches de type ***Apriori*** (par exemple [10, 205, 157, 83]), les approches par ***listes d'occurrences*** (par exemple [222, 223, 18, 148]) et enfin les approches dites par ***projections*** (par exemple [178, 175, 110]) qui sont présentées ci-après.

Il convient d'observer qu'elles ont pour point commun l'utilisation active de la contrainte de support (fréquence). Cette utilisation permet de réduire l'espace des motifs envisagés durant le processus de calcul. Plus précisément, le support est une contrainte ***anti-monotone***. Selon cette propriété, si un motif séquentiel n'est pas fréquent, aucun de ses sur-motifs ne peut être fréquent. Par exemple, si $A \rightarrow B \rightarrow K$ n'est pas fréquent, il n'est pas nécessaire de vérifier si le motif $A \rightarrow B \rightarrow K \rightarrow C$ ou le motif $C \rightarrow A \rightarrow B \rightarrow K$ est fréquent. On peut donc éviter de

considérer l'ensemble de ces motifs lors de la recherche. Cette stratégie est un exemple typique d'utilisation *active* d'une contrainte.

Ces techniques s'appuient sur l'utilisation de structures de données et d'algorithmes dédiés.

3.2.1 Approches de type Apriori

Les approches de type Apriori ont en commun la façon dont l'espace des motifs est exploré. L'exploration se fait en *largeur* (par niveaux), c'est-à-dire que les motifs de taille $k+1$ ne sont considérés qu'après avoir exploré tous les motifs de taille k. De plus, les candidats sont générés en fonction de leur longueur et non de leur préfixe. À ce principe d'exploration se rajoute l'utilisation active de la propriété d'anti-monotonicité du support. Ainsi, une fois les motifs de taille k comptés, des motifs *candidats* de taille $k+1$ sont générés à partir des motifs de taille k qui sont fréquents ; ces motifs candidats étant alors les seuls à être comptés au niveau $k+1$. Ce type d'approche nécessite donc une passe complète pour le comptage des motifs candidats à chaque niveau k.

Agrawal et Srikant ont d'abord étudié le problème de l'exploration des séquences fréquentes [10] et ils ont proposé un algorithme appelé AprioriAll. C'est une amélioration de Apriori [6], une méthode de génération et de test des candidats, assurant que si un candidat peut être fréquent alors il sera généré. Plus tard, ils ont amélioré AprioriAll et ont élaboré un algorithme plus efficace appelé Generalized Sequential Pattern (GSP) [205].

Similaire à la structure de l'algorithme Apriori pour l'extraction de règles d'association, GSP est basé sur la même méthode Générer-Élaguer. La technique utilisée est basée sur une génération plus efficace des candidats, suivie du test de ces candidats pour confirmer leur fréquence dans la base.

Les différentes optimisations apportées aux algorithmes de la famille Apriori utilisent généralement de façon active d'autres contraintes que la contrainte de support, par exemple une contrainte de longueur, qui permet de spécifier le nombre d'éléments composant un motif.

Une autre méthode basée sur le principe Générer-Élagage est Prefix tree for Sequential Pattern (PSP) [157]. La principale différence par rapport à GSP est que les candidats ainsi que les séquences fréquentes sont gérés dans une structure plus efficace. Les méthodes présentées jusqu'ici sont conçues pour dépendre le moins possible de la mémoire principale. Les méthodes présentées par la suite ont besoin soit de charger la base de données, soit de réécrire la base de données, soit de maintenir des pointeurs sur la base de données en mémoire. Ces méthodes peuvent être efficaces lorsque la base de données peut s'insérer en mémoire vive.

Comme il a été souligné auparavant, l'un des principaux problèmes des algorithmes d'extraction est le manque de concentration ou de contrôle de l'utilisateur [169]. Une intéressante famille d'algorithmes, nommée Sequential Pattern mining with Regular Expressions (SPIRIT) [83] adapte les algorithmes basés sur Apriori pour utiliser des expressions régulières (intégrées dans l'algorithme par des automates à états finis) afin de limiter la génération de candidats, ce qui réduit les candidats acceptables pour lesquels un comptage du support est nécessaire.

Étant donné que la tâche de comptage du support d'un candidat est l'opération la plus coûteuse, une autre possibilité est d'éviter l'étape de génération de candidats comme dans les approches par projections.

3.2.2 Approches par listes d'occurences

Les approches de type Apriori sont très consommatrices en accès disque, chaque phase de comptage déclenchant une lecture de toute la base de séquences. Une solution consiste alors à stocker les informations de la base de séquences en mémoire vive, sous la forme de *listes d'occurrences* [224, 222, 223, 18, 148]. Ces listes contiennent, comme leur nom l'indique, les positions des occurrences des différents motifs dans la base de séquences.

Dans [223], les auteurs ont proposé l'algorithme SPADE pour la découverte rapide de séquences fréquentes. SPADE utilise les listes d'occurrences et constitue la base de la méthode d'extraction de motifs séquentiels de longueur variable utilisé dans [123, 124] (voir l'annexe C). L'idée principale dans cet algorithme est un groupement des séquences fréquentes en s'appuyant sur leurs préfixes communs et l'énumération des séquences candidates, grâce à une réécriture de la base de données (chargée en mémoire vive). SPADE a besoin de seulement trois balayages de la base de données afin d'extraire les motifs séquentiels. Le premier balayage vise à trouver les items fréquents, le deuxième à trouver les séquences fréquentes de longueur 2 et le dernier associe aux séquences fréquentes de longueur 2, une table des identificateurs des séquences et des identificateurs des ensembles d'items correspondants dans la base de données (par exemple les séquences de données contenant la séquence fréquente et la date correspondante). Sur la base de cette représentation en mémoire vive, le support des séquences candidates de taille k est le résultat des opérations de jointure sur les tables liées aux séquences fréquentes de taille $(k-1)$ capables de produire ce candidat (ainsi, chaque opération après la découverte des séquences fréquentes ayant la longueur 2 est faite dans la mémoire). L'exploration de la recherche peut se faire en largeur (niveau par niveau) ou en profondeur (branche par branche). L'algorithme cSPADE étend SPADE pour utiliser plusieurs contraintes [222].

Sequential PAttern Mining (SPAM) [7] est une autre méthode qui représente la base de données dans la mémoire principale sous forme de listes d'occurences de vecteurs de bits. Il a été le premier algorithme qui a utilisé une représentation bitmap dans ce domaine.

3.2.3 Approches par projections

Les méthodes de recherche par projections sont plus efficaces pour l'extraction de motifs séquentiels. Elles adoptent un principe "Diviser pour régner" (Divide et impera) et ont pour objectif de réduire les coûts dûs au comptage du support des motifs candidats et de réduire la phase de génération des candidats. Deux idées sont alors mises en avant : (1) réduire la taille de la base de données et (2) éviter d'envisager des motifs n'existant pas dans la base.

L'extraction de motifs séquentiels vise à projeter de manière récursive les séquences de données dans des bases de données plus petites. La solution proposée réside dans le concept de *base projetée* (un sous-ensemble d'une base initiale). L'utilisation de telles bases permet d'accélérer le comptage car la taille des bases projetées est réduite par rapport à la taille de la base initiale, chaque base étant plus facile à traiter.

Proposé dans [110], FREquEnt pattern-projected Sequential PAtterN mining (FreeSpan) est le premier algorithme qui considère la méthode de projection pour extraire des motifs séquentiels. Il trouve premièrement des itemsets fréquents et utilise ceux-ci pour construire des motifs séquentiels.

PREFIX projected Sequential PAtterN mining (PrefixSpan) est un algorithme efficace pour l'extraction de séquences fréquentes [178] qui s'appuie sur le même principe de bases projetées. Il fonctionne de manière récursive en réduisant l'espace de recherche à chaque étape, en

évitant la génération de séquences non-fréquentes. PrefixSpan extrait des séquences fréquentes par une génération de bases de données intermédiaires au lieu de l'approche traditionnelle de génération de séquences de candidats. La projection choisie est réalisée selon le préfixe des motifs à découvrir. PrefixSpan s'avère efficace seulement si une quantité suffisante de mémoire est disponible. Très différent de GSP, PrefixSpan découvre des séquences fréquentes en projetant des bases de données et en comptant le support des items. Cela implique que seuls les supports des séquences qui se produisent réellement dans la base de données sont comptés. En revanche, une séquence candidate générée par GSP peut ne pas apparaître du tout dans la base de données. Le temps pour générer une telle séquence candidate et vérifier si un tel candidat est une sous-séquence de la base des séquences est perdu. Ce facteur contribue à l'efficacité de PrefixSpan par rapport à GSP.

À partir des items fréquents de la base de données, PrefixSpan génère des bases de données projetées qui contiennent les suffixes des séquences de données de la base de données originale, suivant le préfixe (c'est-à-dire l'item fréquent lors de la première projection) utilisé pour la projection. Le processus est répété de manière récursive jusqu'à ce que aucun item fréquent ne se trouve dans la base de données projetée. À chaque fois qu'un item fréquent est découvert dans la base projetée, il est associé en tant que suffixe au préfixe de projection : un nouveau motif séquentiel fréquent est trouvé. Le coût le plus important de PrefixSpan est la génération de bases de données projetées. Pour chaque séquence fréquente découverte, une base de données projetée doit être calculée. Par conséquent, le nombre de bases de données intermédiaires est très important s'il y a beaucoup de séquences fréquentes. Si la base de données est grande, alors PrefixSpan nécessite une quantité importante de mémoire.

Avec le développement de la méthodologie Pattern Growth (PG), même les expressions régulières peuvent être utilisées de manière aisée pour contraindre les processus de fouille de données [175]. Seules les séquences qui satisfont potentiellement la contrainte sont générées. Les séquences qui sont des préfixes sont étendues pour les séquences acceptées.

3.2.4 Recherche incrémentale de motifs séquentiels

Comme les bases de données évoluent, le problème de la mise à jour de motifs séquentiels sur une période longue devient indispensable, car un grand nombre de nouveaux enregistrements peut être ajouté à une base de données. Afin de refléter l'état actuel de la base de données, dans lequel des motifs séquentiels précédents peuvent devenir sans intérêt et de nouveaux motifs séquentiels peuvent apparaître, de nouvelles approches efficaces ont été proposées. Dans [158] est proposé un algorithme efficace, appelé Incremental Sequence Extraction (ISE), pour calculer les séquences fréquentes dans la base de données mise à jour. ISE minimise les coûts de calcul en réutilisant des informations à partir des séquences fréquentes anciennes, à savoir le support des séquences fréquentes. La principale caractéristique nouvelle de l'ISE est que l'ensemble des séquences candidates à tester est sensiblement réduit.

L'algorithme SPADE a été étendu dans l'algorithme Incremental Sequence Mining (ISM) [173] qui est basé sur la bordure négative. Il se situe dans un cadre où l'on considère qu'il est possible d'obtenir, lors d'une extraction initiale, d'autres connaissances que la liste des séquences fréquentes. Afin de mettre à jour les supports et d'énumérer les séquences fréquentes, ISM retient les «séquences fréquentes maximales» et les «séquences non fréquentes minimales». Il a été conçu pour gérer les mises à jour de la base de données où des transactions nouvelles sont ajoutées aux séquences existantes, ou des séquences entièrement nouvelles sont ajoutées à la base de données.

Knowledge base assisted Incremental Sequential Pattern (KISP) [150] propose également de profiter des connaissances préalablement calculées et génère une base de connaissances de motifs

séquentiels calculés avec diverses valeurs de support. Il se situe dans une démarche interactive dans la mesure où il s'intéresse particulièrement aux variations de support.

3.2.5 Situation actuelle

Le développement des contributions autour des motifs séquentiels est principalement dû à leur capacité d'adaptation à de très nombreux problèmes. Pour faciliter cette adaptation les derniers travaux intègrent de plus en plus de contraintes et offrent plus de souplesse sur la définition de motifs séquentiels.

La définition des motifs séquentiels a été adaptée par certains travaux de recherche. Par exemple, dans [139], a été proposé l'agorithme ApproxMap pour extraire des motifs séquentiels approximatifs. ApproxMap propose d'abord de regrouper les séquences de données en fonction de leurs items. Ensuite, pour chaque cluster, ApproxMap permet l'extraction des motifs séquentiels approximatifs liés à ce cluster.

Aujourd'hui, plusieurs méthodes sont disponibles pour découvrir efficacement des motifs séquentiels en accord avec la définition initiale. Des méthodes spécifiques, inspirées des algorithmes précédents, existent dans un large éventail de domaines. Néanmoins, les méthodes existantes doivent être réexaminées parce que les données traitées sont beaucoup plus complexes. L'exploration de flux de données représente une nouvelle classe d'applications où les données entrent et sortent de façon très rapide, voire en temps réel [84, 37]. Afin d'accroître l'utilité immédiate des motifs séquentiels, il est très important d'envisager beaucoup plus d'informations. Ainsi, en associant des motifs séquentiels avec une catégorie de clients ou une information multidimensionnelle, l'objectif principal de l'extraction de motifs multi-dimensionnels séquentiels est de fournir à l'utilisateur final des motifs plus utiles.

Depuis qu'ils ont été définis en 1995 [10], les motifs séquentiels ont reçu beaucoup d'attention. Les travaux sur ce thème sont axés sur l'amélioration de l'efficacité des algorithmes, par de nouvelles structures, de nouvelles représentations ou par la gestion de la base de données dans la mémoire principale. Des extensions ont été proposées en prenant en compte des contraintes associées à des applications concrètes. Dernièrement, motivés par l'utilisation que l'on peut faire de ces motifs, de nouveaus travaux étendent la problématique initiale, notamment à la prise en compte de diverses contraintes ou à d'autres types de motifs.

3.3 Extraction de motifs séquentiels fréquents sous contraintes

La masse de motifs fréquents extraits étant souvent trop importante, elle ne peut être exploitée directement et noie les motifs les plus pertinents pour l'utilisateur parmi ceux trop généraux ou triviaux. D'autre part, l'usage des motifs fréquents est limité. Ils ne permettent pas, par exemple, de découvrir des exceptions.

La fouille de motifs séquentiels fréquents présente les aspects suivants : une entrée généralement très volumineuse, un espace de recherche exponentiel, et un ensemble de solutions trop grand. Cette situation est préjudiciable pour deux raisons. Tout d'abord, les performances peuvent se dégrader : l'exploration devient généralement inefficace voire irréalisable. Deuxièmement, l'identification des fragments de connaissances intéressants, estompés au sein d'une énorme quantité de motifs pour la plupart inutiles, est difficile [201].

Donc, dans la fouille de motifs séquentiels il y a deux difficultés majeures : (1) *l'efficacité* - l'extraction peut retourner un nombre énorme de motifs, dont un nombre important pourrait

être inintéressant pour les utilisateurs et (2) *l'efficience* - il faut souvent du temps et de l'espace de calcul importants pour l'extraction de l'ensemble complet des motifs séquentiels dans une grande base de séquences. Par conséquent, le paradigme de l'extraction de données à base de contraintes a été introduit [204, 222, 175, 26]. La fouille de données en employant des contraintes peut surmonter ces difficultés étant donné que les contraintes représentent généralement l'intérêt de l'utilisateur, ce qui permet de limiter les motifs trouvés à un sous-ensemble particulier satisfaisant certaines conditions fortes. En fait, la contrainte d'extraction introduit les connaissances du domaine dans l'extraction de motifs. En outre, si les contraintes peuvent être poussées en profondeur dans le processus d'extraction de motifs, il est probable d'atteindre l'efficience, puisque la recherche peut être plus concentrée, et dans certains cas, de rendre le processus faisable. La contrainte constitue donc une dimension essentielle de l'extraction de motifs. Ces aspects motivent l'étude de la fouille de motifs séquentiels en utilisant des contraintes. On a vu également l'importance de contraintes dans la recherche de motifs locaux (Chapitre 2) : ils permettent l'extraction à des niveaux de très basse fréquence, où les motifs locaux se trouvent, et en même temps, ils guident la recherche vers des motifs intéressants [24].

Afin de pouvoir manipuler d'une façon compréhensible l'introduction du concept de contraintes et leurs propriétés et influences sur le processus d'extraction de motifs, on introduit les définitions suivantes :

Définition 3.13. ***(langage)*** Le langage L est un ensemble de motifs.

On rappelle qu'un motif traduit une propriété ou un extrait de la base de données. Il décrit un comportement ou rend compte d'un phénomène. Le langage L peut être infini dans certains cas, comme pour les séquences. En effet, pour un ensemble d'items spécifiés E, le langage des séquences L_S regroupe tous les multi-ensembles possibles de L_E. On peut compléter le langage avec une structure en le munissant d'une relation de spécialisation/généralisation, comme proposé par Mitchell dans [164]. Une telle relation structure le langage L et est utile pour localiser les motifs potentiels à extraire et parcourir le moins possible de motifs du langage. Pour les ensembles d'items, l'inclusion $\subseteq$ constitue une relation de spécialisation. Par exemple, comme $A \subseteq AB$, A est plus général que AB et AB est une des spécialisations de A. Similairement, pour les séquences, $\alpha =< \alpha_1\alpha_2...\alpha_n >$ est plus général que $\beta =< \beta_1\beta_2...\beta_m >$ (dénoté par $\alpha \preccurlyeq_S \beta$) si α est une sous-séquence de β.

L'espace de recherche dépend intimement du langage des motifs à extraire et son organisation découle de la relation de spécialisation du langage.

La répartition des motifs séquentiels constitue un triangle ouvert (hypertreillis) car le langage est infini. Contrairement à L_S, l'espace de recherche des séquences présentes dans une base de données réelle, tout en restant un hypertreillis, est fini (voir par exemple la section 3.1.2 ou 5.1 pour des motifs séquentiels fréquents groupés).

Définition 3.14. ***(contrainte)*** Une contrainte q est un prédicat booléen défini sur un langage L. La fonction booléenne de q est $f_q : L \rightarrow \{0, 1\}$ avec $f_q(M) = 1$ si le motif M satisfait la contrainte q.

Une contrainte évalue si un motif φ est intéressant ou non. Elle est aussi appelée prédicat ou requête. Le plus souvent, la contrainte dépend de la base de séquences BS (par exemple, la contrainte de fréquence minimale) même si elle n'y fait pas référence explicitement. Abusivement, on écrit $q(\varphi)$ à la place de $q(BS; \varphi)$. Cette notation met en évidence le lien fort que la contrainte établit entre le langage et la base de données. La définition de la contrainte n'exige aucune propriété particulière sur la contrainte. L'extraction de motifs d'une base de données BS est la sélection des motifs d'un langage L intéressant au regard d'une contrainte q. Plus formellement, il s'agit de déterminer la théorie correspondante.

Définition 3.15. ***(théorie)*** Pour un langage L, une base de données BS et une contrainte q, la théorie $Th(L; BS; q)$ est l'ensemble des motifs de L satisfaisant la contrainte q dans BS.

L'approche de la découverte de motifs contraints a été largement acceptée par la communauté de la fouille de données, car elle donne à l'utilisateur la possibilité de contrôler le processus d'exploration, en introduisant ses connaissances du domaine d'application dans le processus d'extraction et par le rétrécissement du domaine des motifs découverts. L'utilisation des contraintes permet de réduire l'espace de recherche, ce qui contribue de manière significative à atteindre de meilleurs niveaux de performance et de passage à l'échelle [206, 175, 83]. Les résultats obtenus sont plus pertinents par rapport aux besoins de l'utilisateur, et leur nombre réduit évite de saturer sa capacité d'analyse et d'interprétation.

Par la gestion des attentes, on veut dire que les résultats du processus doivent être en conformité avec les attentes des utilisateurs [14]. Cette gestion se fait en contraignant le processus de découverte en utilisant ses connaissances du domaine. Quelques auteurs considèrent que la restriction du domaine de recherche peut transformer le processus d'exploration dans une simple tâche de vérification d'hypothèses [104].

Les contraintes peuvent être examinées et caractérisées de différents points de vue. Dans la suite, elle sont présentées du point de vue de leur application puis du point de vue technique visant leur intégration au sein du processus d'extraction de motifs.

3.3.1 Catégories majeures de contraintes

Du point de vue applicatif, on présente huit catégories de contraintes principalement utilisées dans la littérature sur la base de leur sémantique et forme, en précisant leurs principales caractéristiques [175, 177, 180, 64].

1. une contrainte sur les *items* spécifie les items qui doivent apparaître ou non dans les motifs. En imposant que seuls quelques articles sont d'intérêt, on permet la réduction des motifs découverts. Des exemples de telles contraintes sont des expressions booléennes sur la présence ou l'absence d'items [206] (contrainte *d'inclusion* et contrainte *d'exclusion*).
2. une contrainte de *longueur* spécifie la longueur, exacte, maximale ou minimale des motifs.
3. une contrainte de *largeur* spécifie le nombre exact, maximal ou minimal d'items qui peuvent composer les événements formant les motifs.
4. une contrainte dite *basée sur modèle* est une contrainte qui cherche des motifs qui sont des sous-motifs ou des sur-motifs d'un motif donné (modèle).
5. une contrainte *d'agrégat* dans les applications où les items peuvent être associés à des valeurs. Ce type de contrainte porte sur un agrégat d'items dont la fonction d'agrégation peut être par exemple la somme minimale ou maximale (pour des items à valeurs positives), le maximum, le minimum ou la moyenne. Une contrainte d'agrégat évalue la qualité d'un motif au regard d'une mesure d'intérêt. La forme caractéristique de ces contraintes est $m(X)\theta seuil$ où m est une fonction d'agrégat et $\theta \in \{<, \leq, =, \geq, >\}$. Souvent le réglage du seuil modifie la sélectivité de la contrainte et influence la qualité des motifs associés. Introduite dans [9], la plus utilisée contrainte d'agrégat est certainement la contrainte de fréquence minimale.
6. une contrainte *d'expression régulière* spécifie la forme syntaxique des motifs pouvant être formés à partir des items en utilisant des opérateurs tels que la disjonction ou la fermeture de Kleene. Un motif satisfait ce type de contrainte s'il est accepté par un automate à états finis correspondant à l'expression régulière.

7. une contrainte de *durée* spécifie la durée minimale ou maximale entre le premier et le dernier événement des occurrences du motif séquentiel.
8. une contrainte d'*écart (gap)* spécifie la durée minimale ou maximale entre deux événements consécutifs du motif séquentiel. Elle consiste à imposer une limite à la distance entre deux éléments consécutifs de la séquence. Cette contrainte simple est très utile pour tenir compte de l'impact d'un certain itemset sur un autre, en particulier, quand l'opération survient à un moment donné de temps. De cette manière, il est possible de spécifier qu'un événement a plus d'impact sur les événements proches que sur les événements éloignés.

Les algorithmes d'extraction de motifs séquentiels montrent un niveau acceptable de performance. Toutefois, en présence d'ensembles de données denses ou de seuils de support très faibles, leurs performances sont dégradées. L'utilisation des contraintes et de conjonctions de contraintes peut permettre d'agir d'une manière efficace et d'améliorer les performances de ces processus. De manière efficace signifie, dans le cas de contraintes anti-monotones ou partiellement anti-monotones, que l'extraction de motifs séquentiels avec contraintes peut être réalisée en moins de temps que l'extraction de motifs séquentiels sans contraintes et qu'un nombre réduit de motifs peuvent être présentés à l'utilisateur.

De façon plus générale, les catégories de contraintes précédemment évoquées peuvent être regroupées dans les trois classes suivantes [159] :

- Les contraintes *syntaxiques* regroupent : les contraintes sur les items, sur la longueur, sur la largeur, les contraintes basées sur modèles, les contraintes d'expression régulière. Ces contraintes s'appliquent sur les motifs eux-mêmes. Leur prise en compte est donc essentiellement faite au niveau de l'étape de génération des motifs candidats. Les contraintes syntaxiques formalisent principalement la connaissance de l'expert sur les données [181] pour que les motifs extraits soient compatibles avec ses connaissances. Par ailleurs, intégrer la connaissance de l'expert peut focaliser la fouille sur des informations inattendues (en excluant les motifs triviaux) et peut faciliter ainsi la découverte de connaissances nouvelles [215].
- Les contraintes *temporelles* rassemblent les contraintes de durée et de gap (dénommées dans [180] et [64] comme étant "support-related"). Ces contraintes s'appliquent sur les occurrences des motifs. Leur prise en compte s'effectue au niveau de l'étape de comptage des motifs candidats et elle nécessite l'examen de la base de séquences. Pour d'autres contraintes, savoir si la contrainte est satisfaite peut être déterminé par les motifs fréquents eux-mêmes et leur forme sans faire référence au processus de comptage de support.
- Les contraintes *d'agrégat* : tout comme pour les contraintes syntaxiques, elles s'appliquent sur les motifs eux-mêmes.

3.3.2 Gestion des contraintes

Dans le cadre de la fouille de motifs séquentiels, Srikant et Agrawal [205] ont généralisé l'extraction des motifs séquentiels [10] pour inclure des contraintes de temps, une fenêtre temporelle glissante et une taxonomie définie par l'utilisateur. Garofalakis et al. [83] ont proposé des expressions régulières comme des contraintes pour la fouille de motifs séquentiels et ils ont mis au point une famille d'algorithmes SPIRIT. Les algorithmes utilisent des contraintes relaxées avec des propriétés attractives (comme l'anti-monotonicité) pour filtrer certains motifs/candidats peu prometteurs. L'algorithme c-Spade [222] (qui est une extension de SPADE [223] traitant la contrainte de fréquence minimale) permet d'appliquer un certain nombre de contraintes. L'algorithme explore l'espace des séquences soit niveau par niveau, soit en profondeur. Les contraintes utilisées sont des contraintes de longueur, de largeur, de durée et d'écart.

Du point de vue de l'interaction avec le processus d'extraction de motifs, les contraintes sont généralement caractérisées selon les propriétés de *monotonie*, d'*anti-monotonie* et de concision (*succinteness*) [169, 18, 176]. L'utilisation active d'une contrainte munie d'une de ces propriétés permet de réduire l'espace de recherche en évitant d'explorer les sous-espaces qui ne peuvent pas contenir de motifs satisfaisant la contrainte.

Une contrainte est *anti-monotone* par rapport à la spécialisation si pour tout motif la satisfaisant, tous les sous-motifs qu'il contient satisfont également la contrainte. Une contrainte est *monotone* si pour tout motif la satisfaisant, tous les sur-motifs le contenant satisfont également la contrainte. Une contrainte est dite *succincte* si la spécification qui la définit (une formule précise) permet de générer directement tous les motifs la satisfaisant.

Une mesure monotone (ou anti-monotone) par rapport à la spécialisation est en fait une simple fonction croissante (ou décroissante) par rapport à la spécialisation. Les contraintes monotones/anti-monotones sont bien adaptées à l'extraction de motifs. Tout d'abord, elles peuvent être aisément combinées par conjonction ou disjonction. La classe des contraintes anti-monotones (ou monotones) est stable pour ces opérations. En revanche, la contraposée d'une contrainte monotone (resp. anti-monotone) est une contrainte anti-monotone (resp. monotone). La non-satisfaction d'une contrainte monotone ou anti-monotone donne directement sa condition d'élagage :

Condition d'élagage Si un motif φ ne satisfait pas la contrainte monotone (resp. anti-monotone) q, alors toutes les généralisations (resp. spécialisations) de φ ne satisfont pas la contrainte q.

Si un motif φ vérifie une contrainte anti-monotone q_{am}, tout motif plus général que φ la vérifie également. On peut donc caractériser $Th(B, L, q_{am})$ par l'ensemble de ses motifs les plus spécifiques. Cet ensemble est appelé la frontière positive de la théorie. De manière duale, on peut également définir une frontière négative contenant les motifs les plus généraux qui ne satisfont pas la contrainte. Cette caractérisation a été introduite en extraction de connaissances par Mannila et Toivonen [156]. Ainsi, la contrainte anti-monotone est la contrainte la plus intéressante pour l'extraction de motifs séquentiels. Elle assure un très bon rendement au processus grâce à son action d'élagage.

Une caractérisation complète (selon les propriétés de monotonie, d'anti-monotonie, et de "succinteness") de l'ensemble des contraintes possibles est faite dans [175]. Cette caractérisation fait apparaître que les contraintes d'expression régulière et les contraintes complexes d'agrégat ne sont ni anti-monotones, ni monotones, ni succinctes. Un nouveau cadre, appelé PG, est construit en s'appuyant sur une propriété de préfixe-monotonie par Pei et al. [175, 64] et permet leur prise en compte de façon active lors de l'extraction de motifs dans les bases de séquences. Toutes les contraintes monotones et anti-monotones, ainsi que les contraintes d'expression régulière, sont préfixe monotones. En outre, certaines contraintes d'agrégat fortes, telles que celles impliquant la somme générale ou la moyenne, peuvent également être poussées en profondeur dans un processus d'extraction de motifs de type PG [180].

Il est possible de combiner plusieurs contraintes à la fois. On obtient alors un résultat d'autant plus concis, et ce pour une consommation de ressources d'autant plus réduite (en mode actif). Les **combinaisons** sont importantes pour l'utilisateur car elles enrichissent encore l'expressivité des motifs extraits. Si une contrainte s'avère insuffisante pour exprimer la nature des motifs recherchés, l'utilisateur peut alors la compléter par un ou plusieurs autres critères afin d'affiner ses attentes. Une combinaison de contraintes permet ainsi d'associer leur sémantique respective. En particulier, une **conjonction de contraintes** extrait des motifs satisfaisant la sémantique individuelle de chaque contrainte. En plus de cibler des informations intéressantes,

cette conjonction de contraintes réduit le nombre de motifs extraits et facilite ainsi leur analyse ultérieure. En fait, l'introduction de plusieurs contraintes simultanées implique la vérification de chaque motif potentiel par plusieurs filtres différents (de contenu, temporels et existentiels). Afin de faire face efficacement à cette agrégation des contraintes, l'algorithme doit éviter le test multiple de chaque motif potentiel [62].

Du point de vue technique, une contrainte peut être gérée de façon passive ou de façon active. Le mode **passif** comme dans la stratégie «générer et tester», consiste par exemple à générer tous les motifs fréquents puis à les filtrer (c.-à d. ne retenir que ceux qui satisfont une certaine contrainte). Si l'on considère la consommation des ressources en temps et mémoire, le filtrage est peu rentable, la phase de *post-traitement* pour sélectionner les motifs satisfaisant la contrainte additionnelle ne faisant qu'ajouter une consommation supplémentaire en ressources de calcul. Au contraire, le mode **actif**, en intégrant au plus tôt la prise en compte de la contrainte lors de l'extraction des motifs, permet de concentrer au plus vite les efforts de calcul sur les motifs susceptibles de satisfaire la contrainte et de réduire ainsi les ressources de calcul nécessaires [175, 64, 180, 25]. Pour y parvenir il faut que le prédicat p de la contrainte soit «poussé» pendant l'extraction pour élaguer des portions de l'espace de recherche et réduire le nombre de motifs extraits et le temps de calcul. En utilisant la contrainte, pour que l'extraction soit complète il faut être sûr qu'on découvre tous les motifs qui satisfont le prédicat. Cela signifie qu'à chaque fois que l'on décide de ne pas explorer une partie de l'ensemble L (on dira qu'on élague L), il faut pouvoir prouver qu'aucun motif φ satisfaisant le prédicat p ne s'y trouve (dans ce cas on dira que l'élagage est sûr). Autrement dit, on cherche à n'explorer qu'une partie P de L en s'assurant que la théorie $Th(B, L, p)$ est incluse dans P.

L'approche de **relaxation** approxime la contrainte considérée par d'autres qui possèdent de bonnes propriétés de monotonie ou anti-monotonie. Les motifs satisfaisant ces dernières peuvent facilement être extraits, puis filtrés pour retrouver les motifs satisfaisant la contrainte originelle. On souhaite approximer la théorie de la contrainte originale q par une collection de motifs plus large correspondant à la théorie d'une contrainte plus lâche $q' : Th(L, B, q) \subseteq Th(L, B, q')$. La contrainte moins restrictive q' induite de q, est appelée une relaxation et satisfait l'implication $q \Rightarrow q'$ [203]. L'idée clé est d'obtenir une relaxation vérifiant une propriété de monotonie ou anti-monotonie dans le but de pouvoir réutiliser les algorithmes usuels. À partir de ces théories, un simple filtrage sélectionne alors les motifs satisfaisant q. Une telle approche est une méthode d'optimisation qui préserve la découverte [20] puisque l'élagage issu de la relaxation ne rejette pas de motifs satisfaisant q.

La qualité d'une relaxation diffère en fonction de la taille de sa théorie. Plus précisément, une relaxation est d'autant plus efficace que sa théorie est proche de celle de la contrainte originale. La relaxation (soit monotone, soit anti-monotone) qui approxime au mieux la contrainte originale, est dite optimale (voir un exemple dans la section 5.4).

L'utilisation des contraintes est l'une des principales questions dans l'extraction de motifs séquentiels. D'une part, les motifs découverts correspondent à ceux attendus et d'autre part, les délais de traitement sont plus réduits pour les processus avec des contraintes.

Malgré ces résultats, il est indéniable que l'utilisation de contraintes comme des filtres augmente la complexité de l'extraction de motifs séquentiels. En effet, le temps passé à vérifier l'acceptabilité de chaque motif n'est pas négligeable.

Conclusion

L'état de l'art de la description spatio-temporelle des STIS permet une construction concentrique sur ce sujet. En première ligne, dans le chapitre 1, est présentée une vision d'ensemble sur le processus d'ECD, le domaine général des méthodes utilisées, ses principales étapes et les types principaux de fouilles de données spatio-temporelles.

Puis, dans le chapitre 2, l'exposé se concentre sur des aspects caractéristiques de l'analyse de STIS. Un premier aspect est lié au niveau de l'entité étudiée - pixel ou objet. Puis, on discute la nature supervisée ou non-supervisée de la démarche, les méthodes utilisées pour répondre aux diverses tâches (détection de changements, clustering ou classification) et la nature et complexité de leurs résultats (motifs ou modèles). Pour finir, sont analysés deux types spécifiques de fouille de donnes : la fouille d'information dans les images et la fouille de trajectoires.

Le chapitre 3 focalise la présentation sur les fondements théoriques de l'extraction de motifs séquentiels fréquents, MSF, à partir des STIS introduite par [123] et [124], et les types d'algorithmes d'extraction dédiés. À la fin du chapitre sont exposées l'introduction d'autres contraintes dans le processus d'extraction, les classes de contrainte et la problématique adjacente à la nécessité d'exprimer des contraintes adéquates aux caractéristiques de la base de séquences de la STIS observée et qui répondent à l'intérêt d'utilisateur.

La partie I de cette thèse contient les informations préliminaires nécessaires pour entamer une approche propre d'extraction des motifs séquentiels d'une base de séquences d'évolutions construite avec les données d'une STIS.

Le premier principe de ce mémoire est l'accent mis sur de l'évolution radiométrique des entités de la surface terrestre comme élément principal utilisé pour décrire, caractériser et discriminer ces entités. La base de données contiendra des séquences d'évolutions au niveau pixel pour préserver le niveau de résolution native. La démarche proposée sera non-supervisée, sans aucune sélection d'objet ou de classe a priori. Les images seront considérées dans leur intégralité.

On soutient qu'il est possible d'utiliser efficacement des algorithmes d'extraction de motifs séquentiels avec des contraintes appropriées, sur des données séquentielles, pour découvrir des informations pertinentes et intelligibles, en gardant le processus centrée sur l'utilisateur. On espère que l'efficacité du processus d'extraction sera d'autant plus grande si les trois dimensions de données de STIS sont utilisées : radiométrique, temporelle et spatiale. Les dimensions temporelle et radiométrique pouvant être utilisées complètement lors d'une extraction de MSF, il reste à considérer les caractéristiques spatiales des données dans le cadre d'une nouvelle mesure et contrainte. Afin de pouvoir élaguer l'espace de recherche, il est préférable que la nouvelle contrainte soit anti-monotone. Du fait que la contrainte de support est anti-monotone, la nouvelle contrainte pourra être utilisée en combinaison avec la précédente.

Deuxième partie

Extraction de motifs séquentiels fréquents groupés dans les STIS : définitions et mise en œuvre

Introduction

Un premier type de contrainte utilisé dans ce travail a été celui développé sur la fréquence des séquences d'évolutions au niveau de pixel, mis en évidence par l'utilisation d'un seuil de support. L'application de cette contrainte donne aux motifs extraits une certaine représentativité et pertinence. La contrainte de fréquence (support) étant anti-monotone, son implantation active dans le processus de fouille de données permet la réduction de l'espace de recherche et du temps de traitement. Ces réductions ne sont pas toujours suffisantes et satisfaisantes et des contraintes supplémentaires sont nécessaires. En fait, les contraintes reflètent les caractéristiques de la base de données, les connaissances du domaine et les attentes de l'utilisateur. Le problème est de savoir comment utiliser les contraintes pour préciser les connaissances du domaine d'application et les attentes des utilisateurs, et parallèlement, pour assurer de nouvelles réductions du nombre de motifs et du temps de calcul, tout en permettant la découverte de nouvelles connaissances.

Les informations capturées dans les cellules voisines d'un pixel d'intérêt ou les informations sur les motifs qui les couvrent peuvent fournir des données complémentaires utiles pour cette démarche. Ce type d'informations sont des "données du domaine spatial". En dépit de la quantité supplémentaire d'informations disponibles, il y a relativement peu d'efforts pour extraire les informations spatiales capturées dans les images satellitaires [60].

Dans ce mémoire, pour tenir compte de la spatialité des informations, on introduit une mesure de connexité pour les pixels couverts par un même motif. Cette mesure et les contraintes basée sur elle peuvent mettre en évidence la tendance, naturelle ou induite, des pixels d'une STIS de s'organiser en régions. La construction des contraintes basées sur cette connexité permettra de réduire les nombre de motifs séquentiels fréquents par l'élimination de ceux qui sont insuffisamment connexes. L'objectif est d'obtenir des contraintes anti-monotones et de les mettre en œuvre dans des différentes techniques capables d'améliorer les conditions d'extraction des motifs.

Le chapitre 4 introduit les mesures de connexité locale, globale, moyenne et relative au support minimum, des mesures qui utilisent les informations spatiales sur les emplacements des occurrences d'un motif, et les contraintes construites sur la base de ces mesures. La contrainte de connexité moyenne assure une interprétation assez claire pour l'utilisateur, mais elle n'a pas de propriétés de monotonie qui permettraient une utilisation active de celle-ci dans l'extraction des motifs. Au contraire, les contraintes de connexité globale et relative au support minimum sont anti-monotones, permettant une implantation active et efficiente dans la fouille de données, mais leur interprétation est moins naturelle, rendant plus difficile le choix des seuils.

Le chapitre 5 présente la mise en œuvre des techniques utilisant ces contraintes. Ainsi, la première technique utilise la contrainte de connexité moyenne comme post-traitement. Les contraintes de connexité anti-monotones, globales et relatives au support minimum, permettent le développement d'une technique d'intégration active au sein du processus d'extraction. La technique la plus intéressante est basée sur la relaxation de la contrainte de connexité moyenne avec celle de connexité relative au support minimum.

Chapitre 4

Motifs séquentiels fréquents groupés et contraintes de connexité

Sommaire

Dans ce chapitre, de nouveaux types de contraintes et motifs sont définis (contraintes de connexité et motifs séquentiels fréquents groupés), dédiés à l'extraction de groupes de pixels partageant une évolution temporelle commune, couvrant une surface minimale et satisfaisant en moyenne une connexité spatiale minimale. Certaines définitions préliminaires sont redonnées pour définir une STIS comme un ensemble de séquences temporelles d'évolutions de pixels, d'où un type commun de motif d'exploration de données, le motif séquentiel, peut être extrait. Enfin, on introduit des mesures de connexité utilisées pour définir et extraire les motifs séquentiels groupés fréquents.

4.1 Connexité et mesures de connexité

On considère une STIS qui couvre une même zone à différentes dates. Au sein de chaque image, chaque pixel est associé à une valeur, par exemple, l'intensité de la réflectance de la zone géographique qu'elle représente. Ces valeurs de pixel peuvent être transformées en valeurs appartenant à un domaine discret, à l'aide d'étiquettes pour l'encodage des états de pixel. Ces étiquettes peuvent correspondre à des plages obtenues par la quantification de l'image ou aux classes de pixels résultants d'une classification non supervisée (par exemple, le groupement en utilisant les K-moyennes ou basé sur l'algorithme EM)(comme dans [146], voir l'annexe A).

Définition 4.1. ***(étiquette et état de pixel)*** Soit $E = \{i_1, i_2, \ldots, i_s\}$ un ensemble contenant s symboles distincts denommés *étiquettes*, et utilisé pour coder les valeurs associées aux pixels. Un *état de pixel* est une paire (e, t) où $e \in E$ et $t \in \mathbb{N}$, et tel que t est la date d'occurrence de e. La date t est simplement l'étiquette temporelle de l'image à partir de laquelle la valeur e a été obtenue. Pour le cas où le pas temporel est constant, l'état de pixel a la forme (e, t_k) avec $k \in N$ et t_k la date. On peut avoir le pas d'échantillonnage égal à l'unité de temps.

Pour un pixel ayant comme identificateur spatial (sid) ses coordonnées (x, y) et sa succession d'états, on peut construire sa séquence d'évolution.

Définition 4.2. ***(séquence d'évolution du pixel et STIS symbolique)*** Pour un pixel p, la *séquence d'évolution du pixel* p est une paire $((x, y), seq)$, où (x, y) sont les coordonnées de p et seq est un tuple d'états de pixels $seq = \langle (e_1, t_1), (e_2, t_2), ..., (e_n, t_n) \rangle$ contenant les états de p rangés par ordre croissant de leurs dates d'apparition. Une *STIS symbolique* (où STIS selon le contexte) est alors un ensemble de séquences d'évolution de pixels.

L'analyse des motifs séquentiels d'une STIS conduit à une interprétation naturelle de la notion de *support*. Pour un motif α, le *support* de α est simplement une aire, c'est-à-dire, le nombre total de pixels de l'image ayant une évolution contenant α. Ces pixels sont dits "couverts" par α.

Définition 4.3. ***(pixel couvert)*** Un pixel ayant la séquence d'évolution $((x, y), seq)$ est *couvert* par un motif séquentiel α si α a au moins une occurrence dans seq. L'ensemble de coordonnées de pixels couverts par α est dénoté par $cov(\alpha)$. Par définition, $|cov(\alpha)| = support(\alpha)$.

Donc, pour un motif fréquent α, le seuil σ (ou σ_{rel}) peut être interprété comme la surface minimale (ou surface minimale relative) qui doit être couverte par α. Toutefois, un seuil sur l'aire couverte n'est pas suffisant car, la plupart du temps, la partie intéressante dans les images est constituée de pixels formant des régions. Ainsi, on présente un critère supplémentaire, la mesure de la connexité. La recherche de la connexité entre les pixels couverts d'un même motif est inférée par la situation réelle de la scène et par les post-traitements envisagés des motifs extraits, vus

comme bons candidats pour le clustering, la classification ou simplement l'obtention des objets. En effet, les régions avec des pixels connexes peuvent décrire correctement les objets de la couverture terrestre qui sont habituellement homogènes spécialement dans les zones affectées par l'intervention humaine. La mesure de connexité est basée sur la convention de 8 plus proches voisins [74]. La Figure 4.1 présente la localisation de ces pixels voisins contigus (immédiats).

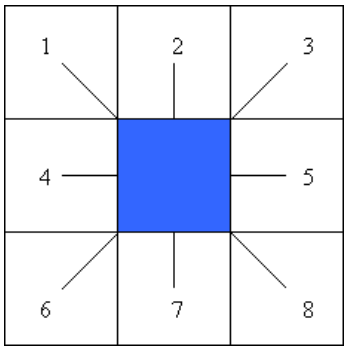


FIG. 4.1 – Les 8 plus proches voisins d'un pixel

À l'aide de cette mesure, on sélectionne les motifs qui couvrent les pixels formant des groupes et qui sont définis comme suit.

Définition 4.4. ***(connexité locale, CL)*** Pour une STIS symbolique $\mathcal{S}$, soit $occ((x,y),\alpha)$ une fonction qui, étant donnés les coordonnées spatiales (x,y) et un motif séquentiel α, indique si α apparaît dans $\mathcal{S}$ à la position (x,y). Plus précisément, $occ((x,y),\alpha)$ est égale à 1 si et seulement si il y a une séquence seq dans $\mathcal{S}$ aux coordonnées (x,y) et α apparaît en $((x,y),seq)$. Autrement, $occ((x,y),\alpha)$ est égal à 0. Si α apparaît en $((x,y),seq)$, alors sa *connexité locale* en (x,y) est [120]

$$CL((x,y),\alpha) = [\sum_{i=-1}^{i=1}\sum_{j=-1}^{j=1} occ((x+i,y+j),\alpha)] - 1 \tag{4.1}$$

La valeur de la $CL((x,y),\alpha)$ est simplement le nombre de pixels dans le 8-voisinage immédiat de (x,y) couvert par α (ou qui supportent α). Il convient de noter que la somme est décrémentée d'une unité un pour ne pas compter l'apparition de α à l'emplacement (x,y).

Définition 4.5. ***(connexité globale, CG)*** La connexité globale d'un motif α est définie par :

$$CG(\alpha) = \sum_{(x,y)\in cov(\alpha)} CL((x,y),\alpha) \tag{4.2}$$

Pour une image, cette mesure donne le nombre de liaisons de tous les pixels, dans un 8-voisinage, qui supportent le motif α.

Définition 4.6. ***(contrainte sur connexité globale)*** Étant donné un motif séquentiel α et un seuil de connexité globale μ_G, la contrainte sur connexité globale est une fonction $f_{CG}(\alpha,\mu_G)$ qui retourne 'VRAI' si $CG(\alpha) \geq \mu_G$, et 'FAUX' autrement.

Théorème 4.1. *La contrainte sur CG est anti-monotone par rapport à la spécialisation.*

Preuve. Puisque le nombre de pixels décroît ou reste constant dans une spécialisation (conformément à la propriété d'anti-monotonie du support), on peut affirmer que pour tout $\alpha \subseteq \beta$ alors $CG(\alpha) \geq CG(\beta)$. Ainsi si $CG(\beta) \geq \mu_G$ alors $CG(\alpha) \geq \mu_G$. Par conséquent $f_{CG}(\beta) = VRAI \Rightarrow f_{CG}(\alpha) = VRAI$ et f_{CG} est donc anti-monotone par rapport à la spécialisation. □

L'inconvénient majeur de la connexité globale (CG) est constitué par les valeurs élevées de CG et par le manque de signification pour l'utilisateur. Néanmoins, sur cette base, on peut construire d'autres mesures de connexité, présentées dans ce qui suit.

La contrainte de connexité introduite peut être considérée une contrainte d'agrégat parce qu'elle évalue la qualité d'un motif au regard d'une mesure d'intérêt dont la détermination implique des calculs. Les calculs pour la mesure ne sont pas faits sur les valeurs des items du motif considéré, mais ils sont liés à l'emplacement spatial des occurrences du motif. De cette manière, on exploite les caractéristiques spatiales des motifs contenus dans la base de données qui est une série d'images dans le temps. Du point de vue des classes de contraintes mentionnées à la sous-section 3.3.1, la contrainte de connexité se rapproche des contraintes temporelles étant appliquée sur les occurrences des motifs. Cependant, l'aspect spatial pris en compte n'est pas inclus explicitement dans l'expression de la séquence et nécessite l'examen des voisinages spatiaux des occurrences d'un motif de la base de données de type images. Ainsi, une nouvelle dimension de données est mise en évidence et prise en compte. En outre, cette contrainte a un évident caractère global, dans le sens que la mesure de connexité est calculée dans toute l'image résultante du motif extrait et pour toutes ses occurrences.

4.2 Contrainte sur connexité moyenne CM et motifs séquentiels fréquents groupés MSFG

Définition 4.7. ***(connexité moyenne, CM)*** La *connexité moyenne* d'un motif α est définie par [120, 119] :

$$CM(\alpha) = \frac{\sum_{(x,y)\in cov(\alpha)} CL((x,y),\alpha)}{|cov(\alpha)|} \tag{4.3}$$

Pour une évolution α, cette mesure donne le nombre moyen de voisins entourant les pixels couverts par α. La définition de la connexité moyenne, comme un rapport entre la connexité globale, CG, et le support d'un motif donné, conduit au problème suivant. Bien que les deux facteurs du rapport puissent, séparément, servir à construire des contraintes anti-monotones à l'aide des seuils minimums, le rapport en lui-même ne peut servir à établir une mesure anti-monotone à base de seuil minimum. Les valeurs des deux fonctions décroissent avec la longueur du motif. La variation du rapport dépend des taux des variations des facteurs du rapport. Si la décroissance relative de la CG est plus petite que celle du support, la connexité moyenne (CM), peut croître avec la longueur du motif et la condition de anti-monotonicité est violée. Une telle situation est rencontrée et présentée au chapitre 6.

Souvent, l'utilisateur peut être intéressé par une combinaison de mesures comme la fréquence et la longueur des motifs. Un tel compromis entre ces deux grandeurs d'intérêt s'exprime par la mesure d'aire : $supp(\alpha) \times longueur(\alpha)$. Ainsi, on exploite une contrainte de support minimum variable selon la taille des motifs [201]. La contrainte de support est envisagée de la façon suivante : plus un motif est long, moins la contrainte qui lui est imposée est restrictive.

Une situation similaire peut être définie dans notre cas. À l'image de la mesure d'aire mentionnée au dessus, en introduisant une contrainte de CM, le produit de type aire (fréquence × connexité moyenne) a une signification précise et concrète :

$$supp(\alpha) \times CM(\alpha) = CG(\alpha) \tag{4.4}$$

Si la mesure de l'aire traduit un compromis entre le support et la longueur, ce produit tenant compte de la fréquence et de la connexité moyenne exprime le nombre de liaisons de tous les pixels

couverts par le même motif. Malheureusement, l'utilisateur ne peut pas avoir une représentation de la valeur immense de CG et, dans ce cas, il ne peut pas fixer de valeur utile pour le seuil en correspondance avec ses attentes d'où l'introduction des mesures relatives de connexité moyenne et connexité relative au support minimum (section 4.3).

Le tableau 4.1 est présenté pour aider l'utilisateur à comprendre l'idée de la connexité moyenne. Pour un carré de côté n pixels, un rectangle avec les côtés n et εn ($\varepsilon \in \mathbb{N}$, l'excentricité) et une chaîne unidimensionnelle de longueur n, les formules de calcul sont données dans ce tableau. Par exemple, pour une connexité moyenne $CM > 6$, on doit avoir la surface couverte par le motif donné plus connexe, plus dense, qu'un carré de côté 7 pixels ou qu'un rectangle de dimensions plus grandes que 5 pixels sur 10 pixels.

Figure compacte	Dimensions	CG	$CM > 6$	$CM > 7$
Carrée	n, n	$8n^2 - 12n + 4$	$n > 7$	$n > 12$
Rectangle	n, εn	$8n^2 - 6n(\varepsilon + 1) + 4$	$\varepsilon = 2, n > 5$	$\varepsilon = 2, n > 9$
Chaîne de pixels	$1, n$	$2n - 2$		

TAB. 4.1 – Connexité globale et moyenne pour des figures géométriques simples

Sur la base de la connexité moyenne, est définie la contrainte correspondante :

Définition 4.8. ***(contrainte sur connexité moyenne, C)*** La contrainte sur connexité moyenne est la fonction $f_{CM}(\alpha, \kappa)$ qui, étant donné un motif séquentiel α et un nombre réel positif κ défini comme *seuil de connexité moyenne*, retourne 'VRAI' si $CM(\alpha) \geq \kappa$, et 'FAUX' autrement.

Enfin, les motifs séquentiels fréquents groupés sont définis comme suit :

Définition 4.9. ***(motifs séquentiels fréquents groupés (MSFG et m-MSFG))*** Soit $\mathcal{S}$ une STIS symbolique, α un motif séquentiel fréquent en $\mathcal{S}$ et κ un seuil de connexité moyenne. Le motif α est appelé un *Motif Séquentiel Fréquent Groupé* (*MSFG*) si $CM(\alpha) \geq \kappa$ dans $\mathcal{S}$. Un MSFG de longueur m est appelé un *m-MSFG*.

Comme on le verra dans la partie III, en pratique, le seuil de support (c'est-à-dire, l'aire couverte minimale) et le seuil de connexité moyenne (c'est-à-dire, degré minimal de regroupement spatiale) permettent la sélection de motifs intéressants pour les applications [120, 126, 119, 160, 116, 121, 122, 118].

4.3 Contrainte sur connexité relative au support minimum CRSM

La mesure de CG a la propriété d'anti-monotonie et peut être utilisée pour la construction d'une contrainte utile pour l'extraction de motifs séquentiels. Les principaux problèmes de cette mesure sont les valeurs très élevées et la difficulté de compréhension de ces valeurs pour l'utilisateur. La connexité moyenne, CM, présente un domaine de valeurs borné (0, 8) et a une signification transparente pour l'utilisateur (le nombre moyen de liaisons ou voisins immédiats d'un pixel), mais l'absence de la propriété d'anti-monotonie empêche son utilisation efficiente. La mesure de connexité suivante peut représenter un bon compromis.

Définition 4.10. ***(connexité relative au support minimum, CRSM)*** La *connexité relative au support minimum* d'un motif séquentiel α est définie comme [121, 122, 118] :

$$CRSM(\alpha) = \frac{\sum_{(x,y) \in cov(\alpha)} CL((x, y), \alpha)}{\sigma} \tag{4.5}$$

Lemme 4.1. *Pour un motif séquentiel fréquent α, on a la relation $CM(\alpha) \leq CRSM(\alpha)$.*

Preuve. Soit α un motif séquentiel fréquent. Selon la Définition 3.8, $support(\alpha) \geq \sigma$ et, selon la Définition 4.3, $support(\alpha) = |cov(\alpha)|$. Ainsi, $|cov(\alpha)| \geq \sigma$. Par conséquent :

$$\frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{|cov(\alpha)|} \leq \frac{\sum_{(x,y)\in cov(\alpha)} LC((x,y),\alpha)}{\sigma} \tag{4.6}$$

Donc, $CM(\alpha) \leq CRSM(\alpha)$. □

Définition 4.11. ***(contrainte sur connexité relative au support minimum, C')*** La contrainte sur connexité relative au support minimum est une fonction $f_{CRSM}(\alpha, \mu)$ qui, étant donné un motif séquentiel α et une valeur de seuil μ, retourne 'VRAI' si $CRSM(\alpha) \geq \mu$, 'FAUX' autrement.

Théorème 4.2. *La contrainte sur CRSM (C') est anti-monotone par rapport à la spécialisation.*

Preuve. Soit α un motif fréquent et σ le seuil de support minimum. Par les définitions 4.5 et 4.8, $CG(\alpha) = \sigma \times CRSM(\alpha)$. Ainsi pour tout $\alpha \subseteq \beta$, puisque le nombre de pixels décroît avec la spécialisation (conformément à la propriété d'anti-monotonie du support), on a $\sigma \times CRSM(\alpha) \geq \sigma \times CRSM(\beta)$ et donc $CRSM(\alpha) \geq CRSM(\beta)$. Ainsi, si $CRSM(\beta) \geq \mu$ alors $CRSM(\alpha) \geq \mu$. Par conséquent $f_{CRSM}(\beta, \mu) = VRAI \Rightarrow f_{CRSM}(\alpha, \mu) = VRAI$ et f_{CRSM} est donc anti-monotone par rapport à la spécialisation. □

Cette contrainte peut être «poussée» profondément dans le processus de fouille de données mais il reste le problème d'une signification claire pour utilisateur.

Supposons que κ soit fixé. Pour un motif α, si $CM(\alpha) \geq \kappa$ alors α satisfait C. Selon la relation 4.6, $CRSM(\alpha) \geq \kappa \times supp(\alpha)/\sigma$. Si l'on souhaite que α satisfasse C', il faut alors poser $\mu = \kappa \times supp(\alpha)/\sigma = \kappa \times \gamma$ où $\gamma = supp(\alpha)/\sigma$ est la sur-couverture du motif.

La relation 4.6 de la lemme 4.1 suggère la possibilité de relaxer CM avec CRSM.

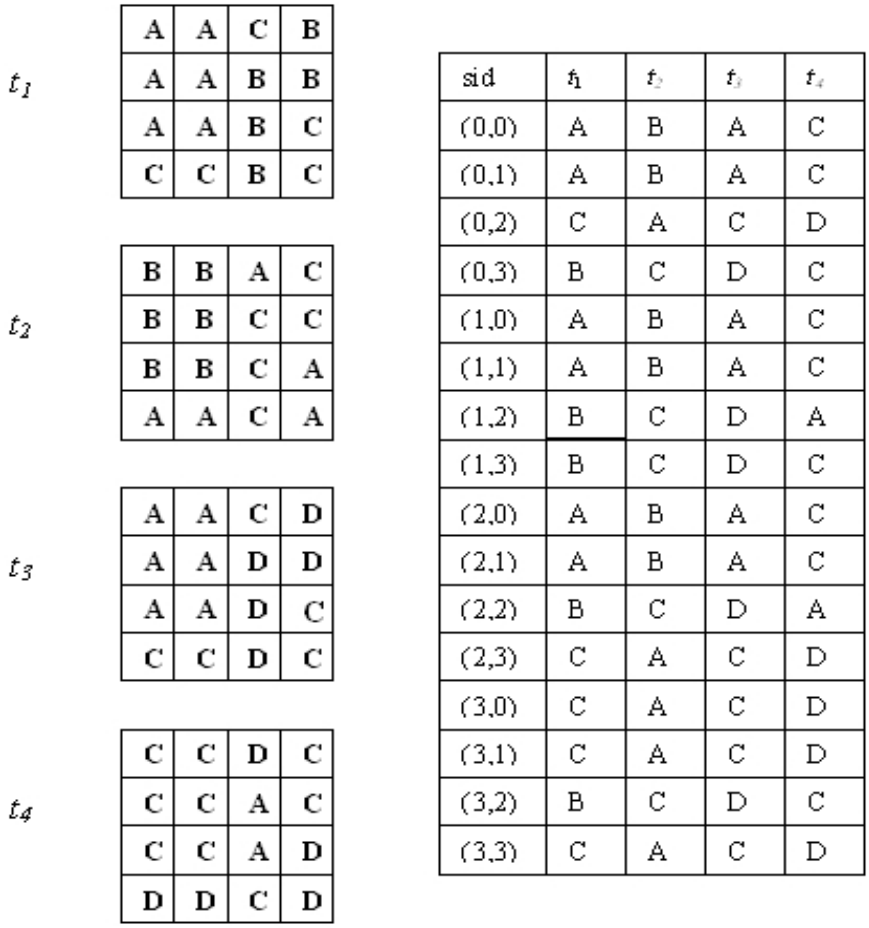
Théorème 4.3. *C peut être relaxée par C'.*

Preuve. Selon la Définition 4.8, la Définition 4.11 et le Lemme 4.1, pour une valeur de seuil donnée κ, $\forall\alpha|\alpha$ est un motif séquentiel fréquent, $C'(\alpha, \kappa) \Rightarrow C(\alpha, \kappa)$. Par conséquent, C peut être relaxée par C'. □

Un jeu de données avec les évolutions temporelles au niveau de pixel accompagné par la base de séquences d'évolutions complètes ainsi obtenue est présenté dans la Figure 4.2. L'ensemble de symboles utilisé a 4 étiquettes (A, B, C et D) et l'identificateur spatial *sid* a la forme (no. de ligne, no. de colonne).

Des exemples de calcul pour les mesures introduites peuvent être réalisés en utilisant le jeu de données de la Figure 4.2 où est présenté également la base de séquences correspondante. On considère une matrice de 16 pixels à 4 instants différents. Les seuils établis sont :

- pour la fréquence (support) $\sigma = 4$
- pour la CM $\kappa = 3$
- pour la CRSM $\mu = 3$



sid	t_1	t_2	t_3	t_4
(0,0)	A	B	A	C
(0,1)	A	B	A	C
(0,2)	C	A	C	D
(0,3)	B	C	D	C
(1,0)	A	B	A	C
(1,1)	A	B	A	C
(1,2)	B	C	D	A
(1,3)	B	C	D	C
(2,0)	A	B	A	C
(2,1)	A	B	A	C
(2,2)	B	C	D	A
(2,3)	C	A	C	D
(3,0)	C	A	C	D
(3,1)	C	A	C	D
(3,2)	B	C	D	C
(3,3)	C	A	C	D

FIG. 4.2 – Jeu de données avec les évolutions des pixels d'une matrice 4×4 au long d'une série de 4 dates et la base de séquences correspondante

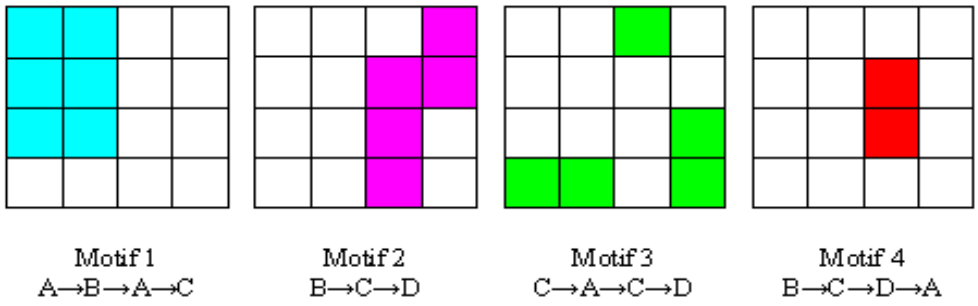


FIG. 4.3 – La localisation de quatre motifs représentatifs de la base de séquences de la Figure 4.2

Les localisations des 4 motifs séquentiels de cette base de séquences sont représentées dans la Figure 4.3. Le calcul des mesures utilisées est présenté ci-dessous :

Motif 1 $A \rightarrow B \rightarrow A \rightarrow C$

$CG(M1) = \sum_{(x,y)\in cov(M1)} CL((x,y),M1) = CL(0,0) + CL(0,1) + CL(1,0) + CL(1,1) + CL(2,0) + CL(2,1) = 3+3+5+5+3+3 = 22$

$Supp(M1) = 6 > \sigma$

$CM(M1) = CG(M1)/supp(M1) = 22/6 = 3,66 > \kappa$

$CRSM(M1) = CG(M1)/\sigma = 22/4 = 5,5 > \mu$

Le motif 1 est un motif séquentiel fréquent groupé, dépassant le seuil de fréquence σ et le seuil de connexité moyenne κ.

Pour le Motif 2, $B \to C \to D$: $supp(M2) = 5 > \sigma$, $CG(M2) = 12$, $CM(M2) = 2,4 < \kappa$, $CRSM(M2) = 3 = \mu$. Le motif est seulement MSF qui accompli la contrainte relâchée.

Le motif 3, $C \to A \to C \to D$: $supp(M3) = 5 > \sigma$, $CG(M3) = 4$, $CM(M3) = 0,8 < \kappa$, $CRSM(M3) = 1 < \mu$. Le MSF n'accomplit aucune contrainte de connexité.

Le motif 4, $B \to C \to D \to A$ est seulement motif séquentiel ($supp(M4) < \sigma, CM(M4) < \kappa, CRSM(M4) < \mu$).

On peut observer qu'un motif incomplet, par exemple le motif 2 $B \to C \to D$, est partialement superposé sur un motif complet (le motif 4 $B \to C \to D \to A$).

En conclusion, dans ce chapitre, différentes mesures de connexité ont été définies et sur leur base on peut construire de contraintes utiles. Dans le cas de la connexité globale, CG, et de la connexité relative au support minimum, CRSM, qui sont anti-monotones pour la spécialisation, les contraintes correspondantes peuvent être «poussées» dans le processus de fouille, en apportant efficience et efficacité. Leur problème est la faible compréhension de leurs valeurs pour l'utilisateur. Au contraire, la mesure de connexité moyenne, CM, est facile à interpréter pour l'utilisateur mais n'a pas de propriété de monotonie. Pour une extraction cohérente et efficiente, une solution peut être la combinaison des contraintes ou la relaxation de la contrainte de CM par une contrainte anti-monotone. Dans le chapitre suivant différentes approches pour utiliser ces contraintes sont présentées.

Chapitre 5

Mise en œuvre des contraintes de connexité

Sommaire

Dans le chapitre antérieur ont été introduites des mesures de connexité qui tirent profit des caractéristiques spatiales de la base de données de type images et qui ont du sens par rapport aux connaissances du domaine. L'objectif algorithmique est d'obtenir des contraintes anti-monotones, pour améliorer les conditions d'extraction des motifs par leur application active dans le processus de fouille de données.

Ce chapitre est dédié à la mise en œuvre des contraintes de connexité. Il présente les techniques développées pour l'application de ces contraintes seules ou en conjonction. Parmi ces techniques, une attention spéciale est accordée à la relaxation de la contrainte de connexité moyenne par la contrainte anti-monotone correspondante.

Le prototype utilisé est SPAtio-TemPorAl Mining (SPATPAM) qui est une évolution de Data Mining Tool 4 Sequential Patterns (DMT4SP) [192]. SPATPAM [63] a été développé dans le cadre du projet Extraction et Fusion d'Informations pour la mesure de Déplacements par Imagerie Radar (EFIDIR) pour extraire des motifs spatio-temporels à partir des STIS et il est en libre accès sur la page web du projet [79].

Le prototype DMT4SP est un outil en ligne de commande pour extraire des motifs séquentiels fréquents, des épisodes et des règles d'épisodes à partir d'une seule séquence ou de plusieurs séquences d'événements. Le support peut être exprimé en termes de nombre d'occurrences ou de nombre de séquences dans lequel le motif apparaît. Le prototype peut inclure des contraintes variées comme des contraintes syntaxiques (longueur, préfixe, suffixe) et temporelles (de durée et d'écart), celles-ci pouvant être combinées lors d'une même exécution.

SPATPAM inclut également la possibilité d'utiliser un critère spatial par la prise en compte des contraintes de connexité proposées dans le chapitre 4. Il est accompagné de routines de prétraitement et post-traitement spécifiques adaptées aux données d'entrée qui sont représentées par des images. Pour les formats de sortie des résultats, il y a la possibilité d'afficher les localisations spatio-temporelles des motifs dans les données (images).

L'algorithme sur lequel s'appuie le prototype utilise une stratégie d'exploration en profondeur et s'inscrit dans le cadre des approches Pattern Growth (PG) [179]. Ainsi la base de séquences est récursivement projetée dans un ensemble de bases plus petites et les motifs séquentiels sont développés dans chaque base projetée en utilisant seulement des fragments localement fréquents. Aucune séquence candidat inutile n'est générée par l'algorithme. Seuls les motifs fréquents sont découverts. En ce qui concerne l'identification des items fréquents à chaque projection, la recherche des occurrences est effectuée dans une base projetée. Une projection virtuelle est utilisée pour réduire le nombre et la dimension des bases projetées, en employant des pointeurs pour garder l'identificateur de la séquence et la position de commencement du suffixe projeté dans celle-ci. Cette technique de projection virtuelle évite la copie physique des suffixes. Elle réduit substantiellement le coût de la projection quand la base projetée peut être contenue dans la mémoire principale.

5.1 Le diagramme connexité - support

Le diagramme de la Figure 5.1 présente les liaisons entre les types de connexités définis et utilisés, et les domaines de motifs séquentiels en fonction de la valeur du support [118]. Le diagramme a deux axes verticaux, la valeur de la Connexité Globale, à gauche, et la valeur de la Connexité Relative au Support Minimum, à droite. L'axe horizontal est la valeur du support des motifs extraits. Sur lui, sont indiqués le seuil de support σ et sa valeur maximale $|BS|$, qui pour le cas de la STIS est le nombre de pixels d'une image. Le diagramme est réalisé pour des seuils de support σ et de connexité moyenne κ donnés. Pour ces seuils, les zones bleues du diagramme

sont interdites. Pour un motif séquentiel α, la contrainte de support ($supp(\alpha) \geq \sigma$) interdit la zone bleue de la bande verticale de gauche et la contrainte de connexité sur CM interdit la bande horizontale bleue inférieure ($CM \geq \kappa$). Le triangle bleu avec le sommet A représente l'impossibilité de la mesure CG à dépasser la valeur du produit $\kappa_M \times supp(\alpha)$, (relation 4.4). La contrainte sur CRSM interdit une bande horizontale inférieure de hauteur μ (égale ou plus grande que celle interdite par la contrainte sur CM, en raison du Lemme 4.1). Le domaine des MSF extraits avec le seuil σ est compris dans le trapèze ayant comme côtés les limites verticales de support $supp = \sigma$ et $supp = |BS|$, l'axe des abscisses et la droite $CG = \kappa_M \times supp$, où κ_M est la valeur maximale du seuil de la connexité moyenne, c.-à-d. 8. Le domaine des MSFG extraits pour le seuil κ donné, la zone blanche du trapèze ABDE, est compris entre les mêmes limites verticales et les droites $CG = \kappa \times supp$ et $CG = \kappa_M \times supp$. En fait, la droite correspondante au seuil $\kappa = 0$ est l'axe des abscisses. Le domaine des motifs extraits avec la contrainte sur CRSM et son seuil $\mu = \kappa$, le trapèze ABGE, est borné par les mêmes limites verticales, la droite $CG = \kappa_M \times supp$ et la droite horizontale $\mu = \kappa$, c.-à-d. la zone blanche et la zone verte.

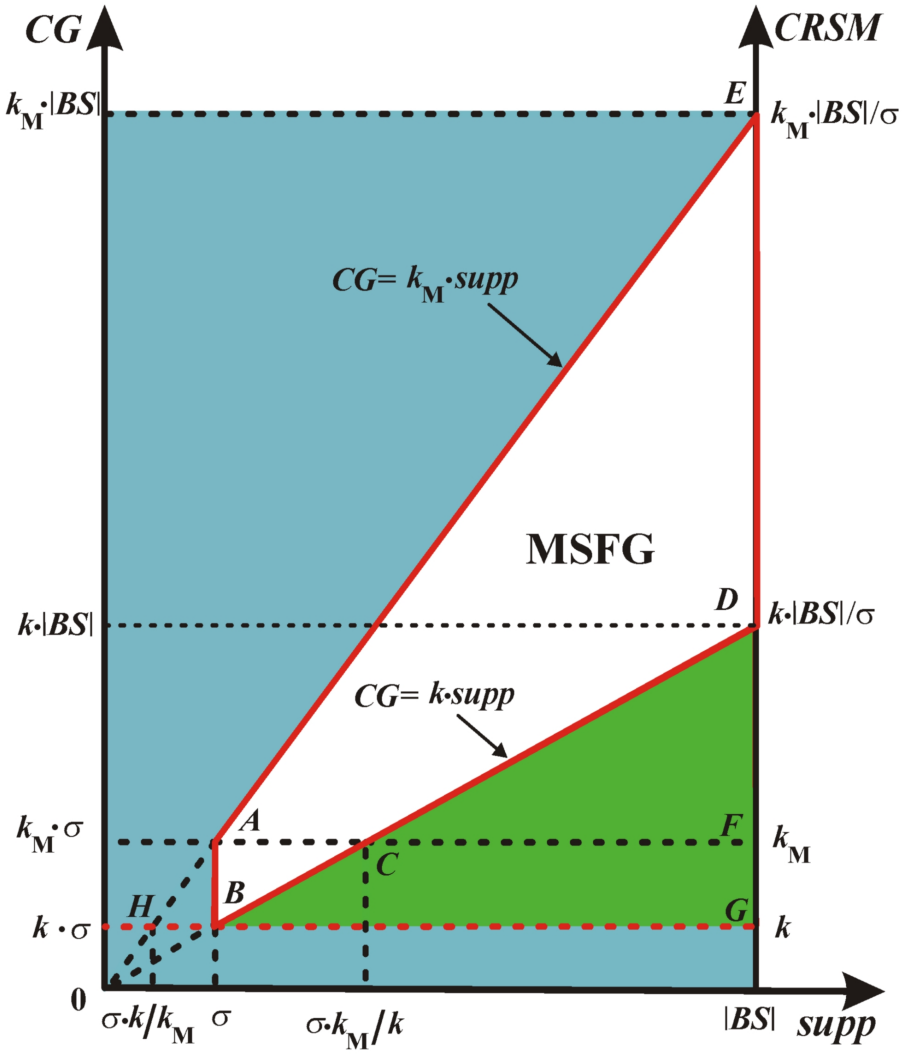


FIG. 5.1 – Le diagramme connexité - fréquence (support) afférente à l'extraction de MSFG

Pour des valeurs différentes du seuil μ, les situations obtenues sont présentées dans la Figure 5.2. Pour $\mu = \kappa$, le domaine des MSFG est inclus dans le domaine de motifs extraits avec la contrainte sur CRSM (Figure 5.2a). C'est la situation dite *optimale* pour une relaxation de

l'extraction des MSFG avec la contrainte sur CRSM, quand la condition de complétude est verifiée et le domaine avec CRSM a l'étendue minimale. Pour des valeurs de seuil μ supérieures au seuil κ, la situation d'incomplétude par rapport à la contrainte sur CM est atteinte. Les MSFG correspondants au seuil κ ne sont pas extraits en totalité avec une extraction commandée avec un seuil de CRSM, $\mu > \kappa$. Une portion du domaine de MSFG reste en dehors comme indiqué dans la Figure 5.2b). Sur le diagramme de la Figure 5.1 on peut voir que pour $\mu = \kappa_M$, le domaine de MSFG du triangle ABC est coupé tandis qu'un domaine parasite du point de vue de MSFG est présent. C'est le domaine correspondent au triangle DCF. Au fur et à mesure que le seuil μ croît, le domaine des motifs extraits avec la contrainte sur CRSM décroît et pour $\mu \geq \kappa \times |BS|/\sigma$, il est inclus dans celui de MSFG (Figure 5.2c).

En effet, soient les seuils σ et κ fixés, et α un MSF ($supp(\alpha) \geq \sigma$). De plus, si α est également MSFG, alors $CM(\alpha) \geq \kappa$. Selon les relations 4.2 et 4.3, l'inégalité devient $CG/supp(\alpha) \geq \kappa$, et avec la relation 4.5 on a $CRSM(\alpha) \times \sigma/supp(\alpha) \geq \kappa$ ou $CRSM(\alpha) \geq k \times supp(\alpha)/\sigma$. Selon la définition 4.11, α satisfait la contrainte C' si $CRSM(\alpha) \geq \mu$. Donc, pour qu'un motif satisfasse simultanément les contraintes C et C' il suffit que $\mu \geq \kappa \times supp(\alpha)/\sigma$. La valeur maximale du membre droit est $\kappa \times |BS|/\sigma$. Ainsi tout motif β ayant $CRSM(\beta) \geq \kappa \times |BS|/\sigma$ satisfait la contrainte sur CRSM et est également MSFG, indépendamment de la valeur de son support possible. Une extraction avec un seuil μ très grand assure des MSFG avec des supports élevés et peut constituer une bonne alternative grâce aux temps réduits de calcul.

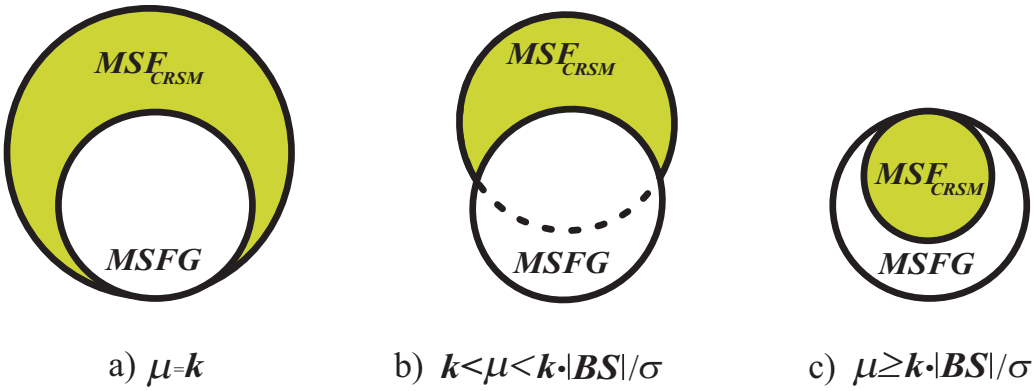


FIG. 5.2 – Les relations d'inclusion des domaines de motifs extraits avec les contraintes sur CM et sur CRSM pour différents valeurs du seuil μ

L'espace de recherche dépend intimement du langage des motifs à extraire et son organisation découle de la relation de spécialisation du langage. La Figure 5.3 illustre l'espace de recherche associé aux langages des motifs séquentiels après une extraction avec une contrainte anti-monotone.

La répartition des motifs séquentiels constitue un triangle ouvert car le langage est infini. Sur la Figure 5.3a), la représentation usuelle associée à l'arbre de préfixes, les motifs les plus généraux (respectivement, spécifiques) sont situés vers le sommet (respectivement, vers l'ouverture du triangle). Contrairement à L_S, l'espace de recherche des motifs séquentiels présents dans une base de données est fini. La zone bleue continue schématise les motifs extraits de la base de données si une contrainte anti-monotone est «poussée» dans le processus d'extraction. Dans le cas d'application d'une contrainte sans propriété de monotonie, l'espace de solutions n'est pas continu. L'objectif des algorithmes d'extraction de motifs est de localiser au mieux les motifs désirés, conformément au prédicat des contraintes appliquées, à travers le vaste espace de recherche afin d'en parcourir le minimum.

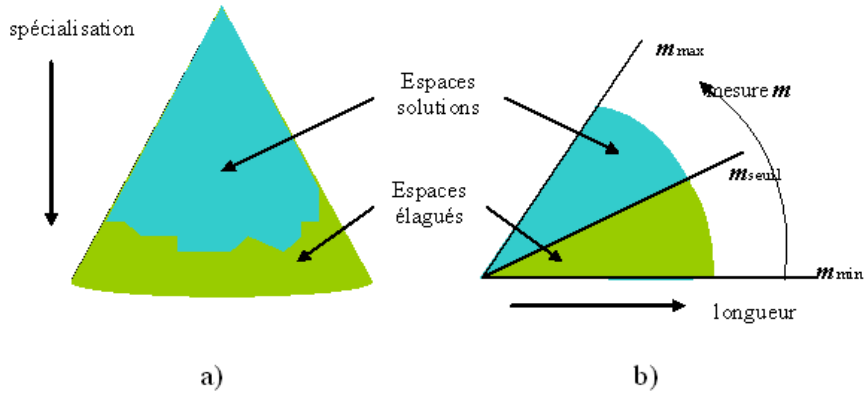


FIG. 5.3 – L'impact de l'application d'une contrainte anti-monotone sur l'espace de solutions des motifs séquentiels. a) représentation usuelle b) représentation en coordonnées polaires

L'espace de recherche peut être organisé également en coordonnées polaires avec ρ correspondant à la longueur des motifs et l'angle φ à la valeur d'une mesure anti-monotone m. Avant l'extraction, le domaine des valeurs de la mesure est compris entre une valeur minimale (m_{min}) et une valeur maximale (m_{max}). L'application de la contrainte avec le prédicat $m \geq m_{seuil}$ coupe la zone verte avec $m_{min} \leq m \leq m_{seuil}$ comme indiqué en Figure 5.3b).

Chaque point de la zone bleue représente un motif extrait sous une contrainte anti-monotone. Tous ses sous-motifs appartiennent à la même zone. Si on veut construire le "trajet" de la croissance de la longueur d'un motif, sa spécialisation, toutes les "étapes" sont dans la zone bleue. Une sortie dans la zone verte, la zone où la contrainte anti-monotone n'est pas satisfaite, est irréversible.

5.2 Application de la contrainte sur connexité moyenne CM (post-traitement)

La première méthode utilisée consiste en l'obtention de tous les motifs fréquents (par application de la contrainte anti-monotone de support) puis leur vérification du point de vue de la contrainte sur CM [120, 119]. Cette contrainte n'a pas de propriété de monotonie et elle peut être utilisée seulement pour le filtrage des motifs antérieurement obtenus. Selon la définition 4.9, les motifs extraits avec cette méthode sont des MSFG. Ce type de post-traitement a une faible efficience, étant très consommateur de temps de calcul. Ces hypothèses sont vérifiées dans la Partie III par les résultats obtenus à partir des jeux de données réels de différentes STIS.

Pour toutes les techniques proposées, on a implanté une fonction qui rend possible l'application des contraintes de connexité pour des données images 2D, l'utilisateur ayant la possibilité de choisir la technique désirée au travers d'un paramètre.

La fonction est appelée pour chaque motif M trouvé fréquent, pour vérifier si le motif satisfait la (les) contrainte(s) de connexité désirée(s). Les valeurs pouvant être retournées sont :

Ⓐ le motif respecte la (les) contrainte(s), il peut être affiché et il doit être spécialisé pour générer d'autres motifs nouveaux possibles (le parcours en profondeur continue)

Ⓑ le motif ne respecte pas la (les) contrainte(s), il ne doit pas être affiché mais il doit être spécialisé pour générer d'autres motifs nouveaux possibles (le parcours en profondeur conti-

nue)

Ⓒ le motif ne respecte pas la (les) contrainte(s), il ne doit pas être affiché, il ne doit pas être spécialisé pour générer d'autres motifs (le parcours en profondeur s'arrête, un élagage est réalisé) C'est la seule variante intéressante du point de vue de la réduction de l'espace de recherche (contrainte anti-monotone)

La table de valeurs pour le post-traitement avec la contrainte sur CM est présentée dans le Tableau 5.1.

CM	
V	F
A	B

TAB. 5.1 – Table de valeurs de sortie de la fonction pour la contrainte sur CM où Vrai et Faux sont des réponses pour la vérification de la contrainte (V pour $CM(M) \geq \kappa$, F pour $CM(M) < \kappa$)

L'organisation de la fonction de vérification de la contrainte sur CM est décrite par la Fonction 1.

Fonction 1 vérification de la contrainte sur CM

Entrée κ - seuil de connexité moyenne, M - un motif déjà identifié comme fréquent, $supp(M)$ - le support de M, les localisations des occurrences de M

Sortie la fonction retourne une valeur (Ⓐ ou Ⓑ) selon que le motif respecte la (les) contrainte(s) posée(s) et caractérise l'étape suivante de l'extraction de motifs fréquents

1: construction d'une image contenant les localisations des occurrences du motif M (pixels couverts)
2: calcul de la connexité locale pour chaque pixel couvert en parcourant l'image
3: calcul de la connexité globale $CG(M)$ comme la somme des connexités locales des pixels couverts
4: $CM(M) \leftarrow CG(M)/supp(M)$ // calcul de la CM pour le motif M
5: **if** $CM(M) \geq \kappa$ **then**
6: **return** Ⓐ
7: **else**
8: **return** Ⓑ
9: **end if**

Dans le diagramme connexité - support de la Figure 5.1, le domaine des MSFG extraits avec cette méthode est le trapèze blanc ABDE.

Une discussion plus détaillée sur les dépendances des paramètres d'entrée et sur les caractéristiques des résultats obtenus avec ces méthodes d'extraction est présentée dans la Partie III qui détaille les applications.

5.3 Application de la contrainte sur connexité relative au support minimum CRSM (poussée)

Disposant des contraintes anti-monotones, sur CG et CRSM, on peut les «pousser» dans le processus de fouille de données. La propriété d'anti-monotonie de CRSM, implique que

$f_{CRSM}(M') = FAUX \Rightarrow f_{CRSM}(M) = FAUX$

pour tous les sur-motifs M de M'. Toutes les fois qu'un motif viole la contrainte on peut élaguer le motif sans perte de complétude. Dans cette section on présente l'implantation active de la contrainte sur CRSM. La table de valeurs de la fonction correspondante est donnée dans le Tableau 5.2.

CRSM	
V	F
A	C

TAB. 5.2 – Table de valeurs de sortie de la fonction pour la contrainte sur CRSM où Vrai et Faux sont des réponses pour la vérification de la contrainte (V pour $CRSM(M) \geq \mu$, F pour $CRSM(M) < \mu$)

Le module de vérification de la contrainte sur CRSM est décrit dans la Fonction 2.

Fonction 2 vérification de la contrainte sur CRSM

Entrée σ - seuil de fréquence minimale, μ - seuil de connexité relative au support minimum, M - un motif déjà identifié comme fréquent, $supp(M)$ - le support de M, les localisations des occurrences de M

Sortie la fonction retourne une valeur (Ⓐ ou Ⓒ) selon que le motif respecte la (les) contrainte(s) posée(s) et caractérise l'étape suivante de l'extraction de motifs fréquents

1: construction d'une image contenant les localisations des occurrences du motif M (pixels couverts)
2: calcul de la connexité locale pour chaque pixel couvert en parcourant l'image
3: calcul de la connexité globale $CG(M)$ comme la somme des connexités locales des pixels couverts
4: $CRSM(M) \leftarrow CG(M)/\sigma$ // calcul de la $CRSM$ pour le motif M
5: **if** $CRSM(M) \geq \mu$ **then**
6: **return** Ⓐ
7: **else**
8: **return** Ⓒ
9: **end if**

Dans le digramme connexité - support de la Figure 5.1, le domaine des MSF extraits avec cette méthode est le trapèze borné par les verticales $supp = \sigma$ et $supp = |BS|$, la droite $CG = \kappa_M \times supp$ et la droite horizontale $\mu = \kappa$. On voit que cette approche permet l'extraction supplémentaire de MSF connexes de point de vue de la mesure de CRSM mais qui ne sont pas groupés de point de vue de la Définition 4.9. Pour les valeurs $\mu = \kappa$, la condition de complétude par rapport à la contrainte sur CM est satisfaite : le domaine de MSFG est compris dans le domaine de motifs extraits avec la contrainte sur CRSM (Figure 5.2a). Au fur et à mesure que μ croît, le nombre de motifs contraints par CRSM diminue et on commence à perdre des MSFG mais également des MSF connexes mentionnés au-dessus (Figure 5.2b). Pour des valeurs élevées du seuil μ, ($\mu \geq \kappa \times |BS|/\sigma$, tous les motifs extraits sont MSFG (Figure 5.2c) et l'extraction de ce type peut être une alternative pour obtenir rapidement les plus connexes et en même temps fréquents motifs. Pour une extraction sous la contrainte $CRSM \geq \mu$ le seuil de connexité moyenne obtenu est

$$\kappa = \mu \times \sigma/supp = \mu/\gamma \tag{5.1}$$

où $\gamma = supp/\sigma$ est la sur-couverture du motif. Une fois comprise la signification de la sur-couverture, l'utilisateur peut accéder à la valeur convenable de μ et tirer profit d'une extraction avec la contrainte sur CRSM «poussée» au sein de l'extraction, cette méthode étant la plus efficiente (temps d'extraction réduits significativement). Les problèmes à résoudre sont les

MSF supplémentaires extraits pour $\mu \leq \kappa \times |BS|/\sigma$ et l'incomplétude pour $\mu > \kappa$. L'étude expérimentale réalisée dans la Partie III valide cette approche.

5.4 Relaxation de la contrainte sur CM par la contrainte sur CRSM ($\mu = \kappa$)

Pour optimiser le processus d'extraction des MSFG, on peut chercher à réécrire la contrainte sur CM sous la forme d'une conjonction de deux contraintes dont l'une est anti-monotone. Il sera alors possible de pousser la contrainte anti-monotone et de profiter de l'élagage qu'elle apporte. Dans la sous-section 3.3.2, on réalise une relaxation de la contrainte sur CM par une contrainte plus lâche sur CRSM. La méthode de relaxation approxime la contrainte considérée sur CM, qui a une signification claire pour l'utilisateur, par une autre sur CRSM, qui possède une propriété d'anti-monotonie [118]. La contrainte moins restrictive sur CRSM, induite sur CM, est appelée une relaxation et satisfait l'implication $CM \Rightarrow CRSM$. L'idée clé est d'obtenir une relaxation vérifiant une propriété d'anti-monotonie dans le but de pouvoir réutiliser les algorithmes usuels. Les motifs satisfaisant la contrainte sur CRSM peuvent facilement et efficacement être extraits, grâce à son pouvoir d'élagage, puis filtrés pour retrouver les motifs satisfaisant la contrainte originelle, celle sur CM. Une telle approche est une méthode d'optimisation de la fouille qui préserve la découverte [20] puisque l'élagage issu de la relaxation ne rejette pas de motifs satisfaisant CM. Pour accomplir la condition de complétude il faut être sûr que $Th(BS, L, CM) \subseteq Th(BS, L, CRSM)$. Dans le diagramme de la Figure 5.1 on peut voir que cette condition est accomplie pour $\mu = \kappa$. Dans notre contexte, la relaxation est une méthode d'optimisation pour rendre faisables et améliorer certaines extractions de contraintes complexes. Elle peut être vue comme une forme de pré-traitement.

Dans la Figure 5.4 sont présentés les espaces de solutions dans le cas de la relaxation. L'application de la contrainte anti-monotone q_{AM} conduit à l'espace des motifs élagués (en vert) et à l'espace continu de solutions (en bleu foncé). Sur cet espace on applique le filtrage de la contrainte q qui n'a aucune propriété de monotonie et les motifs obtenus couvrent des zones qui ne sont pas connexes (en bleu clair).

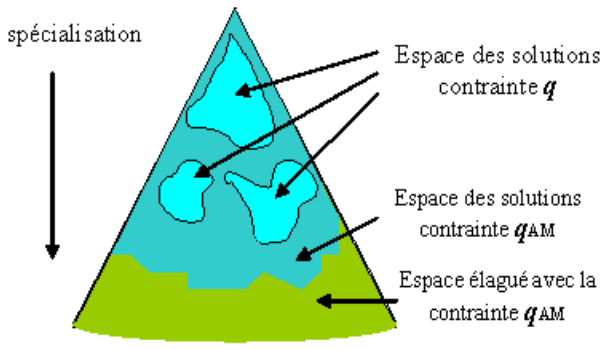


FIG. 5.4 – Les espaces de solutions dans le cas d'une relaxation

La table de valeurs pour la relaxation de la contrainte sur CM par la contrainte anti-monotone sur CRSM est présentée dans le Tableau 5.3. Puisqu'il existe l'implication $CM \Rightarrow CRSM$, la situation $CM = Vrai$ et $CRSM = Faux$ est impossible et c'est pourquoi il n'est pas prévu de réponse pour cette situation dans la table. La mesure de CM n'étant pas anti-monotone, dans

la situation $CM = Faux$ et $CRSM = Vrai$, le motif peut être spécialisé pour la génération d'autres candidats.

		CRSM	
		V	F
CM	V	A	-
	F	B	C

TAB. 5.3 – Table de valeurs de sortie de la fonction pour la méthode de relaxation de la contrainte sur CM par la contrainte sur CRSM

La fonction de vérification de la relaxation de la contrainte sur CM avec la contrainte sur CRSM est présentée dans la Fonction 3.

Fonction 3 vérification de la relaxation de la contrainte sur CM avec la contrainte sur CRSM ($\mu = \kappa$)

Entrée σ - seuil de fréquence minimale, $\mu = \kappa$ (seuil de connexité relative au support minimum = seuil de connexité moyenne), M - un motif déjà identifié comme fréquent, $supp(M)$ - le support de M, les localisations des occurrences de M

Sortie la fonction retourne une valeur (Ⓐ, Ⓑ ou Ⓒ) selon que le motif respecte la (les) contrainte(s) posée(s) et caractérise l'étape suivante de l'extraction de motifs fréquents

1: construction d'une image contenant les localisations des occurrences du motif M (pixels couverts)
2: calcul de la connexité locale pour chaque pixel couvert en parcourant l'image
3: calcul de la connexité globale $CG(M)$ comme la somme des connexités locales des pixels couverts
4: $CM(M) \leftarrow CG(M)/supp(M)$ // calcul de la CM pour le motif M
5: $CRSM(M) \leftarrow CG(M)/\sigma$ // calcul de la $CRSM$ pour le motif M
6: **if** $CM(M) \geq \kappa$ **then**
7: **return** Ⓐ
8: **else**
9: **if** $CRSM(M) \geq \kappa$ **then**
10: **return** Ⓑ
11: **else**
12: **return** Ⓒ
13: **end if**
14: **end if**

L'étude expérimentale sur des données réelles réalisée dans la Partie III prouve la supériorité de cette approche par rapport à celles des sections 5.2 et 5.3. En comparaison avec le post-traitement impliqué par l'application de la contrainte sur CM, l'approche de relaxation offre efficience et efficacité en raison de la réduction de l'espace de recherche. En comparaison avec le fait de pousser la contrainte anti-monotone sur CRSM, l'avantage est la complétude de l'extraction de MSFG.

5.5 Conjonction des contraintes sur CM et CRSM ($\mu > \kappa$)

La conjonction combine les avantages de ces deux contraintes : l'élagage de la contrainte sur CRSM et le filtrage de MSFG par la contrainte sur CM. Cette approche complète le domaine des valeurs du seuil de CRSM ($\mu > \kappa$). Elle n'est pas une relaxation à cause de l'échec de la condition

d'implication des contraintes sur CM et CRSM (ou d'inclusion des théories comme illustré dans la Figure 5.2c, d). La conjonction des contraintes élimine les MSF supplémentaires extraits avec l'approche de la section 5.3 par le filtrage de la contrainte sur CM. Dans le diagramme de la Figure 5.1, on peut voir que la conjonction n'assure pas la complétude du point de vue des MSFG. Parmi les motifs élagués se trouvent une partie de MSFG de support réduit. L'approche est une alternative très efficiente et efficace pour le cas d'extraction des MSFG qui sont en même temps très connexes et fréquents (pour $\mu \geq \kappa|BS|/\sigma$ le filtrage avec la contrainte sur CM n'est plus nécessaire).

Le Tableau 5.4 présente les valeurs de sortie de la fonction pour la conjonction des contraintes sur CRSM et CM avec $\mu > \kappa$.

		CRSM	
		V	F
CM	V	A	C
	F	B	C

TAB. 5.4 – Table de valeurs de sortie de la fonction pour la conjonction des contraintes sur CM et CRSM

Par rapport à la fonction de vérification de la relaxation de la contrainte sur CM avec la contrainte sur CRSM, la fonction utilisée ici demande en entrée deux valeurs distinctes pour les seuils de connexité qui doivent satisfaire la condition ($\mu > \kappa$). Dans la Fonction 4 est décrite la vérification de la conjonction de contraintes sur CRSM et CM ($\mu > \kappa$).

Fonction 4 vérification de la conjonction de contraintes sur CRSM et CM ($\mu > \kappa$)

Entrée σ - seuil de fréquence minimale, μ - seuil de connexité relative au support minimum, κ - seuil de connexité moyenne, M - un motif déjà identifié comme fréquent, $supp(M)$ - le support de M, les localisations des occurrences de M

Sortie la fonction retourne une valeur (Ⓐ, Ⓑ ou Ⓒ) selon que le motif respecte la (les) contrainte(s) posée(s) et caractérise l'étape suivante de l'extraction de motifs fréquents

1: construction d'une image contenant les localisations des occurrences du motif M (pixels couverts)
2: calcul de la connexité locale pour chaque pixel couvert en parcourant l'image
3: calcul de la connexité globale $CG(M)$ comme la somme des connexités locales des pixels couverts
4: $CM(M) \leftarrow CG(M)/supp(M)$ // calcul de la CM pour le motif M
5: $CRSM(M) \leftarrow CG(M)/\sigma$ // calcul de la $CRSM$ pour le motif M
6: **if** $CRSM(M) \geq \mu$ **then**
7: **if** $CM(M) \geq \kappa$ **then**
8: **return** Ⓐ
9: **else**
10: **return** Ⓑ
11: **end if**
12: **else**
13: **return** Ⓒ
14: **end if**

L'application à l'extraction de bases de séquences réelles de STIS atteste de l'efficience de cette approche pour les motifs très fréquents et connexes.

Conclusion

Selon les conclusions de la Partie I, des mesures de caractéristiques spatiales sont introduites et sur cette base, on vise à construire des contraintes anti-monotones susceptibles d'être implémentées profondément dans le processus d'extraction de motifs séquentiels. La mesure générique introduite est la connexité des pixels couverts par le même motif, qui exploite la propriété de dépendance spatiale des données de type motif séquentiel. Pour déterminer le degré de proximité, il faut utiliser des relations spatiales non explicitement stockées dans les bases de données.

Ainsi, dans le chapitre 4, les informations spatiales des images d'occurrence des motifs séquentiels de STIS sont utilisées dans le processus d'extraction. Différentes mesures de connexité des pixels couverts par un même motif sont définies. L'ordre logique d'introduction est le suivant : 1) la connexité locale, CL, définie pour un pixel, qui établit le nombre des pixels voisins conformes du point de vue du motif couvrant ; 2) la connexité globale, CG, qui totalise les connexité locale (CL) de tous les pixels couverts par le même motif ; 3) la connexité moyenne qui divise la CG par le support du motif et 4) la connexité relative au support minimum, CRSM, qui est le rapport entre la CG et le seuil de fréquence, σ. Sur cette base, des contraintes ont été développées pour filtrer ou pour être «poussées» profondément dans le processus de la fouille de données séquentielles. Sur la base de la contrainte de connexité moyenne, qui a une signification claire pour l'utilisateur, on décrit le nouveau concept de motif séquentiel fréquent groupé (MSFG). N'ayant pas de propriété d'anti-monotonie, cette contrainte peut être utilisée seulement pour le filtrage des MSF extraits par application active de la contrainte de support. Au contraire, les contraintes sur la connexité globale, CG, et sur la connexité relative au support minimum, CRSM, sont anti-monotones et peuvent être «poussées» dans le processus d'extraction. Cette application active permet de bénéficier de ses propriétés opérationnelles, la réduction de l'espace de recherche et implicitement la réduction du temps d'exécution, et de l'adéquation avec l'intérêt de l'utilisateur.

Puisque les valeurs de seuil de ces contraintes anti-monotones ne présentent pas de signification assez claire pour l'utilisateur des conjonctions de contraintes sont adoptées dans le chapitre 5 pour tirer profit de la combinaison de leurs effets opérationnels. Un diagramme de la connexité globale en fonction du support met en évidence les caractéristiques des différentes approches d'extraction avec des contraintes. Les approches discutées sont un filtrage (post-traitement) avec la contrainte sur CM, une intégration de la contrainte sur CRSM au sein du processus d'extraction, une approche par relaxation de la contrainte sur CM par la contrainte sur CRSM (obtenue pour la relation $\mu = \kappa$ entre les seuils correspondantes) et une conjonction des contraintes sur CM et CRSM ($\mu > \kappa$). Les meilleurs résultats peuvent être obtenus avec la relaxation optimale de la contrainte sur CM par la contrainte sur CRSM. Le chapitre 5 contient les tables de valeurs de sortie et les organisations des fonctions de vérification pour chacun des quatre régimes de fonctionnement.

La littérature insiste sur le fait qu'un processus d'ECD de qualité requiert une interactivité

et une itérativité fortes avec l'utilisateur. L'interactivité du processus doit mettre en avant l'utilisateur au sein de l'extraction. Le processus doit pouvoir accepter des contraintes variées afin de couvrir l'attente d'utilisateur. Le cas de la contrainte de connexité est pertinent. Dans le chapitre 4 sont proposés différentes types de cette contrainte. Le chapitre 5 présente différentes modalités d'utilisation. Ainsi, l'utilisateur peut choisir entre l'implantation comme filtre d'une contrainte sans propriétés de monotonie, l'intégration au cœur du processus d'une contrainte anti-monotone, la relaxation par une contrainte anti-monotone plus lâche et la combinaison de contraintes.

La validation des hypothèses faites dans cette Partie II et dans la partie I est obtenue dans la Partie III où les méthodes d'extractions envisagées sont appliquées sur différents types de données de STIS.

Troisième partie

Extraction de motifs séquentiels groupés fréquents dans des STIS : applications et résultats

Introduction

Dans le but d'évaluer la pertinence des concepts introduits en Partie II, et afin de tester les méthodes d'extraction correspondantes, sont réalisées des expériences sur différents types de STIS. En effet, comme rapporté par les auteurs de [60], l'information portée par les STIS dépend de ses caractéristiques :

- spectrales (longueur d'onde ou fréquence, propriétés réfléchissantes ou émissives) ;
- spatiales (angle du capteur, forme et taille de l'objet, position, site, distribution, texture) ;
- temporelles (changements en temps et en position) ;
- polarimétriques (effets d'objet à l'égard des conditions de polarisation de l'émetteur et du récepteur).

Les applications décrites dans la Partie III couvrent tous ces quatre types de caractéristiques.

La démarche est non-supervisée et permet l'analyse de la totalité des données fournies par les images satellitaires. Pour permettre le traitement des informations contenues dans les données satellitaires au niveau de la résolution native, l'analyse est faite au niveau pixel. Pour réduire le domaine de valeurs des pixels et se ramener à une description symbolique, on utilise une discrétisation des valeurs initiales par intervalles non superposés établis en utilisant des équirépartitions ou des discrétisations élaborées par l'expert. Les séquences temporelles des valeurs de pixels donnent leurs évolutions. Ces sont les informations de base d'une STIS qui assurent une caractérisation spatio-temporelle de la scène observée. Au travers de nos méthodes, les caractérisations les plus singulières sont exhibées à partir des motifs extraits et des cartes de visualisation spatio-temporelles de ces derniers.

Pour extraire les évolutions d'intérêt pour l'utilisateur, on applique des contraintes. Leur rôle est de réduire l'espace de recherche et d'implémenter les connaissances du domaine selon l'intérêt de l'utilisateur. La première contrainte utilisée est celle de support (ou fréquence) qui assure une pertinence aux motifs extraits. Par application de cette contrainte on obtient les motifs séquentiels fréquents, MSF, i.e. les motifs couvrant une surface minimum. La deuxième contrainte appliquée est nouvelle et tire profit des informations spatiales. C'est la contrainte de connexité basée sur les différentes mesures de connexité développées en Partie II. Ainsi, la contrainte basée sur la connexité moyenne, facilement interprétable par l'utilisateur, assure l'extraction d'un nouveau type de motifs, les motifs séquentiels fréquents groupés, MSFG. En raison de l'absence de propriétés de monotonie, la contrainte de connexité moyenne ne peut fonctionner que comme un filtrage, avec pour conséquence négative l'augmentation du temps d'exécution. Les contraintes anti-monotones de connexité développées (la contrainte sur la connexité globale, CG, et la contrainte sur la connexité relative au support minimum, CRSM) peuvent être poussées dans le processus de fouille de données et, par l'élagage assuré, peuvent réduire efficacement l'espace de recherche et par conséquent le temps d'exécution. Sur la base de ces contraintes sont réalisées des extractions avec la relaxation de la contrainte sur CM par la contrainte sur CRSM, (le cas $\mu = \kappa$), et avec une conjonction de contraintes (pour $\mu > \kappa$).

Les meilleurs résultats sont obtenus avec la relaxation de la contrainte sur CM. Pour tous ces types d'extractions développés, on a réalisé une étude quantitative des influences des paramètres d'extraction comme le nombre de symboles, s, les seuils pour le support, σ, et connexité, μ_G, κ, et μ, et la longueur des motifs extraites, L. Les paramètres de sortie suivis sont les différents rendements et fonctions de transfert d'extraction, décrits par le nombre de motifs extraits et visités et par le temps d'exécution. L'étude est complétée par des stratégies de sélection des motifs et l'interprétation de leur qualité du point de vue de la couverture de la scène et de la pureté de description.

L'application de cette approche à différents types de données démontre la généricité du concept de MSFG. On utilise différentes STIS couvrant soit la zone de Fundulea, Roumanie (projet ADAM) [117, 125, 120, 119, 121, 122], soit le lac Mead, Etats-Unis [120, 119], soit la zone Chamonix Mont Blanc, France (projet EFIDIR) [126, 116]. La première STIS fournit des données optiques. Les autres STIS sont obtenues par interférométrie et polarimétrie radar. L'extraction de MSFG se révèle être utile pour l'identification et la surveillance des objets et phénomènes, des cultures agricoles, d'autres types de couverture terrestre, des déformations de la croûte terrestre ou des évolutions de mécanismes dominants de rétrodiffusion décrits pour des localisations spatio-temporelles précises.

Chapitre 6

Données optiques : la STIS du projet ADAM (Fundulea, Roumanie)

Sommaire

La première STIS utilisée dans ce travail est celle du projet Assimilation de Données par Agro Modélisation (ADAM) [46]. L'assimilation de données inclut des techniques qui, par l'association des Données avec des Modèles, permettent l'estimation des paramètres et des variables d'état d'un système au cours du temps [144, 141]. Ce projet a répondu à la demande d'appliquer l'assimilation des données de télédétection dans les modèles de fonctionnement des cultures [92], visant à intégrer les mesures physiques issues des images satellitaires aux modèles utilisés en agronomie. Plus précisément, l'objectif principal était d'élaborer une méthodologie visant à assimiler l'imagerie spatiale à haute résolution et la répétitivité temporelle dans des modèles agro-physiologiques couplés à des modèles de transfert radiatif.

Une grande quantité d'images issues des différents capteurs a donc été acquise sur une zone d'agriculture intense de la plaine du Danube en Roumanie, près de Bucarest. Cette STIS est choisie pour son contenu en divers objets de dimensions plus grandes que la résolution du satellite, pour des champs agricoles avec cultures contrôlées, permettant d'avoir une vérité terrain sûre pour évaluation. Son domaine temporel est suffisamment long et sa cadence temporelle est appropriée pour évaluer les cycles phénologiques.

6.1 Données de la STIS ADAM

6.1.1 Les images SPOT

Les données utilisées proviennent de la base de données ADAM (disponible à `http://kalideos.cnes.fr`) qui fournit des séries temporelles d'images des canaux SPOT(1,2)-HRV (Haute Résolution Visible) et SPOT4-HRVIR (Haute Résolution Visible InfraRouge). Cette STIS de haute résolution est composée d'images acquises par SPOT 1, 2, et 4 opérant en mode multispectral. Tous les satellites SPOT [1] évoluent à une altitude approximative de 820 km, sur des orbites quasi polaires, caractérisées par une inclinaison de 98,7° (ce qui permet l'héliosynchronisme). La période de révolution des satellites SPOT est de 101,4 minutes, le cycle orbital a une durée de 26 jours, et la résolution spatiale résultante est de 20 m. De plus, le système SPOT a une capacité de dépointage de +/- 31,06° fait qui leur confère une répétitivité d'acquisition de 1 - 3 jours.

La durée de l'acquisition prise en compte dans la STIS ADAM s'est étendue sur 286 jours, d'octobre 2000 à juillet 2001. Les images avec beaucoup de nuages ou neige sont retirées de la séquence. Il en résulte une STIS de 20 images irrégulièrement échantillonnées dans le temps. De plus, parce qu'il y a des zones non imagées à certains instants, en raison des variations de l'angle lors de l'acquisition, et pour concentrer l'étude sur de zones avec un spécifique entièrement agricole et bien maîtrisées, les dimensions des images sont choisies 1000×1000 pixels centrés sur la zone Fundulea, et sur les parcelles d'un institut de recherche agricole pour lesquelles une verité terrain est disponible.

Les images sont calibrées avec précision de point de vue radiométrique et géométrique par le Centre National d'Études Spatiales, Toulouse, France, en rendant possible une comparaison spatiale et temporelle entre images. En outre, les images sont corrigées de point de vue atmosphérique avec le code SMAC [190] et les caractéristiques des aérosols sont mesurées avec un photomètre solaire automatisé.

Il existe trois niveaux de pré-traitement des données SPOT [19, 171]. La correction du premier niveau est une correction radiométrique consistant en l'égalisation des sensibilités des détecteurs et en une correction supplémentaire, géométrique, qui utilise la trajectoire du satellite de façon à supprimer l'effet panoramique, la rotation et la courbure de la terre, et la variation d'altitude du

satellite par rapport à l'ellipsoïde de référence. Au deuxième niveau, une correction géométrique supplémentaire est effectuée par projection cartographique de façon à pouvoir combiner l'image avec d'autres informations géographiques et une projection cartographique avec points d'appuis (ou levée GPS sur le terrain) est aussi utilisée. Au troisième niveau, la projection cartographique considérée utilise non seulement des points d'appui, mais aussi un modèle numérique du terrain de façon à éviter les distorsions liées au relief à l'aide d'une triangulation spatiale et d'une interpolation. Les images finales obtenues ont une taille de 3000×2000 pixels, soit une superficie de $60 \times 40 km^2$. Un pixel représentant un site spatialement localisé sur la surface terrestre admet comme zone de variation dans la STIS un disque dont le diamètre est de 1,5 pixels. En d'autres termes, sur toute la séquence des images, un même site se situera dans un cylindre spatio-temporel dont le diamètre spatial est de 1,5 pixels. Par ce traitement, les images finales sont rendues géométriquement superposables.

Le Tableau 6.1 résume les caractéristiques des données utilisées dans ce travail.

Caractéristique	STIS ADAM 2000-2001	STIS utilisée
Nombre d'images	39	20
Domaine spectral du canal 1	500 - 590 nm	500 - 590 nm
Domaine spectral du canal 2	610 - 680 nm	610 - 680 nm
Domaine spectral du canal 3	780 - 890 nm	780 - 890 nm
Canal synthétique 4		IVDN (NDVI)
Intervalle temporel maximal entre deux images	31 jours	39 jours
Intervalle temporel minimal entre deux images	1 jour	1 jour
Intervalle temporel moyen entre deux images	7,8	12,75
Nombre de lignes	3000	1000
Nombre de colonnes	2000	1000
Résolution spatiale	20 m	20 m
Précision du recalage géométrique	cylindre $\approx$ 0.5 pixels	cylindre $\approx$ 0.5 pixels

TAB. 6.1 – Caractéristiques des données

Dans l'annexe A sont présentés les pré-traitements effectués sur les données de la STIS ADAM afin d'améliorer les résultats de la fouille de données.

6.1.2 La scène observée

Le site ADAM ($44°27'38,43''N; 26°37'14,34''E$) correspond à une zone d'agriculture intensive, où les satellites SPOT1, SPOT2 et SPOT4 peuvent se compléter mutuellement pour fournir la fréquence élevée de revisite temporelle requise. La ferme de production de semences de l'Institut National de Recherche et Développement de l'Agriculture (en roumain Institutul Naţional de Cercetare Dezvoltare Agricolă, INCDA) (INRDA) a de vastes champs agricoles dont la superficie oscille entre 15 ha et 40 ha. Hormis le blé, la culture principale, il y a des cultures de maïs, d'herbe du Soudan, d'orge, de petit pois, de soja, de pois chiches, d'avoine, de haricot, de moutarde et de colza. Les autres objets de la scène peuvent être classés dans 'routes', 'eau', 'forêts' et 'villes'. La topographie de cette région est généralement plate (altitude moyenne de 68 m) avec une petite partie de la zone correspondant à des pentes bordant la rivière Mostiştea et à plusieurs micro-dépressions ('crov', en roumain). Une vérité terrain est disponible pour la période 2000-2001 pour les champs qui appartiennent à l'INRDA. Même si elle représente 5,9% de la scène, elle peut pourtant servir à évaluer les résultats. Cette information n'est pas utilisée au sein du processus d'exploration de données lui-même mais seulement dans l'étape

d'évaluation des résultats.

Dans la télédétection, le concept d'identification et de caractérisation des objets de la couverture du sol est basé sur leur comportement spectral (la signature spectrale). Le rayonnement électromagnétique incident, provenant en principal du soleil, est réfléchi, absorbé ou transmis. La proportion relative entre ces processus dépend de chaque matériau. Dans la Figure 6.1 sont présentées les caractéristiques spectrales des trois principales composantes de la couverture du sol : le sol, la végétation (une céréale), et l'eau de la rivière [21]. Le domaine spectral, le visible et le proche infrarouge (PIR), comprend les trois bandes satellitaires de SPOT.

Dans l'annexe D sont décrits succinctement les mécanismes qui influencent les propriétés spectrales de ces principaux objets observables dans la scène de la STIS ADAM. La compréhension de ces mécanismes rend possible la déduction de la nature physique et physiologique des objets terrestres par leurs réponses spectrales.

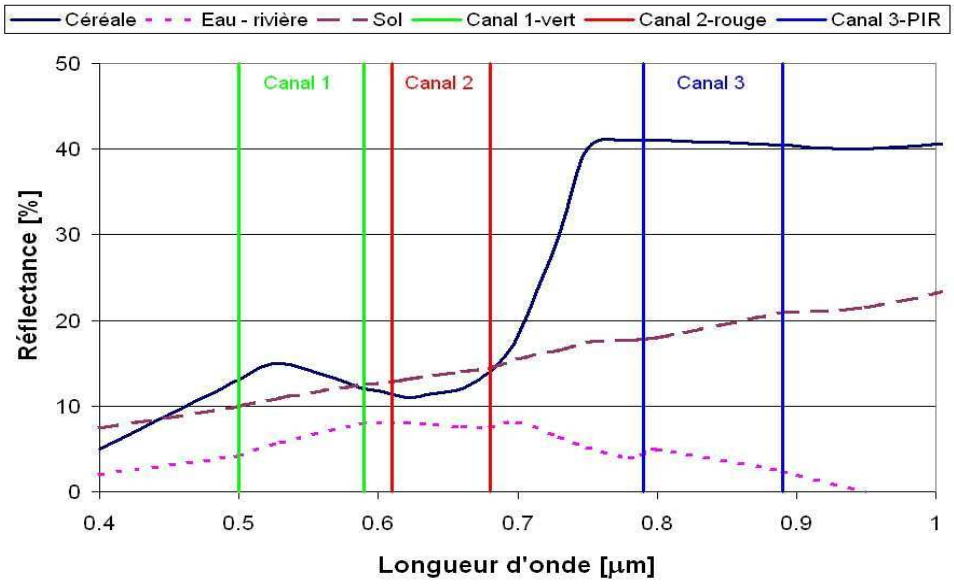


FIG. 6.1 – Les courbes de réflectance spectrale des principales composantes de la thématique de la scène [21].

Quand un objet de la couverture du sol présente des réflexions dans deux ou trois canaux, des indices "spécialisés", très utiles pour interpréter les données satellitaires peuvent être introduits. Les combinaisons des bandes spectrales visible et proche infrarouge permettent de discriminer les surfaces de sol nu ou l'eau de la végétation. Ces combinaisons arithmétiques de bandes sont dénommées «indices de végétation spectrale» [108, 55] et nous fournissent un aperçu spatial des structures de couverture de végétation. Les indices spectraux visent à renforcer la contribution spectrale de la végétation tout en minimisant les contributions de l'arrière-plan du sol, de l'angle du soleil et de l'atmosphère en combinant diverses bandes spectrales dans le visible et proche infrarouge. Ainsi, pour la végétation, parmi la multitude des indices définis, il y a l'Indice de Végétation Différentielle Normalisée (en anglais Normalized Difference Vegetation Index, NDVI) (IVDN) [195, 211, 212] qui est l'indice le plus couramment utilisé. L'indice fournit des méthodes d'estimation de la production primaire nette sur les différents types de biome, d'identification d'écorégions, de surveillance des modèles phénologiques de surface végétative de la terre et d'évaluation de la longueur de la saison de croissance. Cet indice compense largement la modification des conditions d'éclairage, la pente de la surface et les aspects d'angle de visée. Une image IVDN permet d'accroître sensiblement la discrimination de la végétation par rapport

à d'autres types de couverture de la surface terrestre. Il est possible de mieux identifier les zones de végétation malsaine ou stressée, leurs indices étant inférieurs à ceux de la végétation verte saine. Les valeurs IVDN des roches et du sol nu très sec sont petites en raison de leur réflexion semblable dans les deux bandes. Donc, dans une image IVDN les tonalités claires sont associées à une couverture dense de végétation saine. La relation de calcul pour cet indice est :

$$IVDN = \frac{(R_{PIR} - R_R)}{(R_{PIR} + R_R)} \tag{6.1}$$

où R_{PIR} est la réflectance dans le domaine spectrale du proche infrarouge et R_R est la réflectance dans le rouge. Les bandes qui sont utilisées pour le calcul de l'IVDN pour le satellite SPOT HRV sont : la bande 2 pour le rouge, la bande 3 pour le PIR.

6.2 Résultats quantitatifs - Statistique des données et réglage de paramètres

L'objectif principal de cette section est de guider l'utilisateur pour choisir les paramètres d'extraction de motifs séquentiels d'une base de données particulière. Pour une STIS de type ADAM on cherche à obtenir des motifs fréquents, connexes et de longueurs différentes pour une description pertinente des évolutions des objets de la scène. Ce nombre de motifs doit être compatible avec les possibilités d'analyse de l'utilisateur mais supérieur à la diversité thématique de la scène.

Les grandeurs qui constituent les variables dans cette étude sont : le nombre de symboles pour les valeurs des pixels, s, les seuils des mesures anti-monotones de support, σ, de connexité globale μ_G et de connexité relative au support minimum, μ, le seuil de connexité moyenne, κ, la longueur, L et le taux de sélectivité d'extraction. Les fonctions surveillées sont le nombre de motifs extraits, N_m, le nombre des motifs visités pour tests, N_{vis}, le temps d'extraction, t_{ex}, la fonction de transmission du filtre équivalent, FT, le nombre de pixels couverts par un motif, N_C, et le taux d'extraction. Le développement de la démarche passe par les étapes de l'extraction de MS, de MSF, et de MSFG, selon les contraintes utilisées.

Comme objectifs spécifiques, on a l'intention d'obtenir pour une extraction :

1. toutes les longueurs de motifs pour détecter des informations avec divers degrés de généralité ou spécialisation sur l'évolution des valeurs des pixels : objectif $N_m(L)$;
2. une réduction du nombre total de motifs (par divers critères : les plus fréquentes, les plus connexes) offerts à l'utilisateur pour analyse : objectif N_m réduit ;
3. obtention d'un nombre suffisant de motifs longs pour une caractérisation de l'évolution de la valeur des pixels pour tous les objets grands de la scène : objectif L grande ;
4. une efficacité d'extraction raisonnable : objectif t_{ex} petit ;
5. un degré élevé de couverture de la vérité terrain : objectif N_C grand.

6.2.1 Extraction des motifs séquentiels

Conformément à la relation 3.3, le nombre de motifs possibles, N_{mp}, croît exponentiellement avec le nombre d'images. Dans notre cas, $n = 20$, même pour de valeurs réduites du nombre de symboles, s, le nombre de motifs possibles est considérable.

Dans l'annexe A, une première tentative de réduction de l'espace de recherche des motifs a été réalisée par la quantification impliquant l'introduction de s intervalles de valeurs pour les pixels.

Toutes les expériences sur les données de la STIS ADAM sont réalisées sur un ordinateur standard (processeur Intel Core 2 Duo @ 3GHz avec 4 Go de mémoire RAM sous le système d'exploitation Linux noyau 2.6.22.19-02 x86_64).

Le Tableau 6.2 présente les caractéristiques d'extraction des MS contenus dans la base des séquences du projet ADAM et le nombre de motifs possibles pour différentes valeurs du nombre de symboles, s. Pour des raisons de volume de mémoire impliquée, ici sont présentées seulement les situations pour des valeurs réduites du nombre de symboles.

s	Temps d'extraction [s]	Nombre de motifs contenus	Nombre de motifs possibles	Mémoire utilisée [GB]	Densité des motifs séquentiels [%]
2	119,15	510.027	2.097.150	4,14	24,320
3	530,38	10.367.679	5.230.176.600	4,75	0,198
4	1442,24	77.317.194	1.466.015.503.700	17,48	0,005

TAB. 6.2 – L'extraction des motifs séquentiels de la base de séquences du projet ADAM

La densité de motifs séquentiels de la base de séquences ADAM est définie comme le rapport entre le nombre de motifs séquentiels, N_{MS} et le nombre de motifs possibles N_{MP}

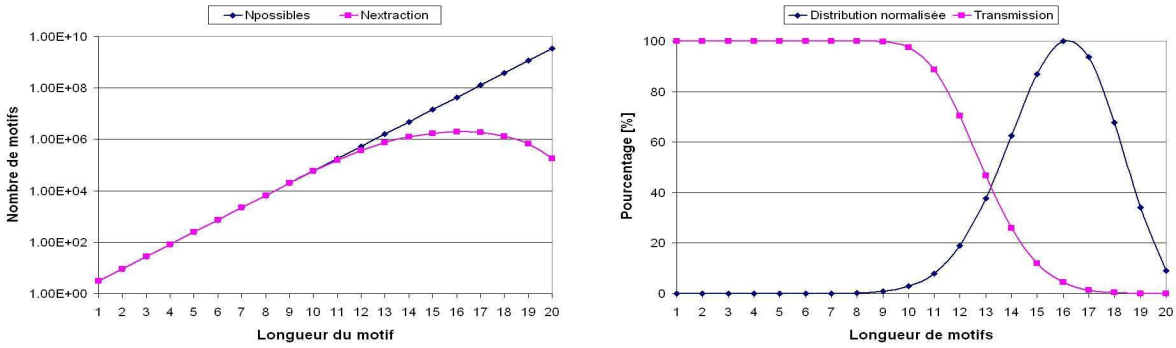
$$\rho_S = \frac{N_{MS}}{N_{MP}} \tag{6.2}$$

Cette densité décroît très fortement avec la croissance du nombre de symboles, s.

L'opération d'extraction des motifs séquentiels peut être équivalente à l'action d'un filtre qui transmet les motifs possibles de la base des séquences vers des motifs séquentiels en écartant des motifs en fonction de leur longueur. Par exemple, pour $s = 3$, les variations, suivant la longueur, du nombre de motifs possibles et du nombre de motifs contenus dans la base des séquences ADAM sont présentées dans la Figure 6.2a).

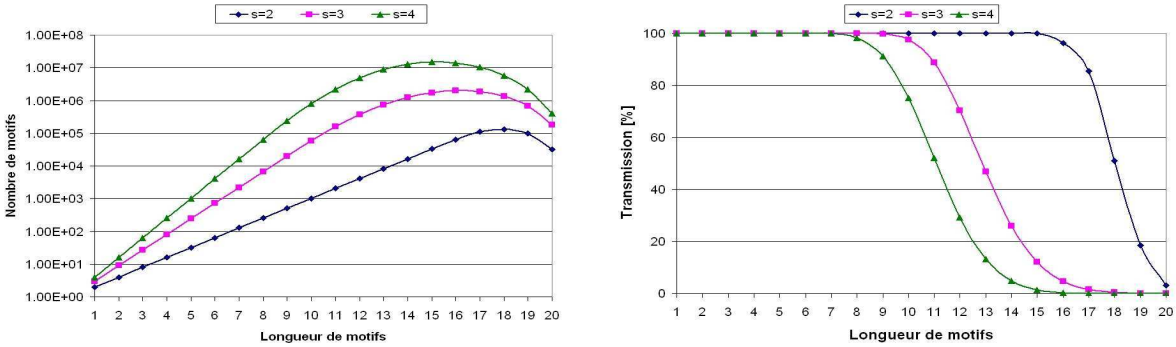
Le rapport entre la grandeur de sortie, le nombre de motifs séquentiels extraits distribués par leurs longueurs, et la grandeur d'entrée, le nombre de motifs possibles, donne la caractéristique de transmission, une mesure du transfert de ce filtre équivalent. La corrélation entre la courbe de transmission et la distribution normalisée de la grandeur de sortie est présentée dans la Figure 6.2b). Le maximum de la distribution des motifs est décalé vers des grandes longueurs par rapport à la pente de coupure de la caractéristique de transmission. Ce fait est dû à la croissance exponentielle de la grandeur d'entrée qui est d'autant plus amplifié que le nombre de symboles croît.

Si la distribution de ces motifs contenus dans la base de séquences suivant leur longueur est étudiée, en ayant comme paramètre le nombre de symboles s, les courbes présentées dans la Figure 6.3a) sont obtenues. Pour des petites longueurs, les portions linéaires (où le comportement exponentiel du nombre de motifs possibles peut être reconnu) correspondent à la situation dans laquelle tous les motifs possibles sont retrouvés et extraits de la base de séquences du projet ADAM. Pour toutes les longueurs, le nombre de motifs séquentiels croît avec le nombre de symboles s. Pour des s petits, les nombres de motifs de longueur maximale, $L = 20$, contenus dans cette base de séquences restent considérables, entre $3,2 \times 10^4$ pour $s = 2$ et $4,11 \times 10^5$ pour $s = 4$.



a) b)

FIG. 6.2 – a) Nombre des motifs possibles et extraits de la base de séquences ADAM suivant leur longueur, pour un nombre de symboles utilisés, $s = 3$ et b) La distribution normalisée des motifs séquentiels extraits et la caractéristique de transmission équivalente suivant la longueur des motifs pour un nombre de symboles utilisés, $s = 3$.



a) b)

FIG. 6.3 – a) Nombre des motifs extraits de la base de séquences ADAM suivant la longueur des motifs et le nombre de symboles utilisés, s et b) Les caractéristiques de transmission équivalentes pour le processus d'extraction des motifs séquentiels.

Dans la Figure 6.3a), les courbes de distribution présentent des maximums situés à des longueurs d'autant plus courtes que le nombre de symboles est grand.

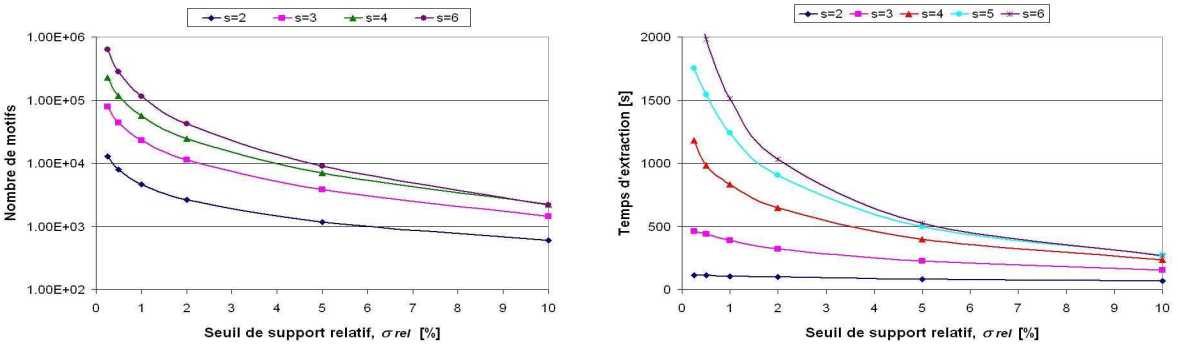
Les caractéristiques des filtres équivalents aux extractions effectuées avec différents nombres de symboles, s, sont présentées dans la Figure 6.3b). Ces courbes de transmission sont caractéristiques pour la base de séquences donnée. Les courbes commencent à couper les motifs possibles vers les longueurs courtes pour un nombre de symboles grand. La comparaison entre les graphiques de la Figure 6.3 a) et b) montre que l'écart observé entre les pentes des courbes de transmission et de distribution de motifs par longueur croît avec le nombre de symboles utilisés, s.

En dépit des réductions significatives du nombre des motifs extraits en comparaison du nombre de motifs possibles, les nombres de motifs mis à la disposition de l'utilisateur restent considérables et il est préférable de les réduire.

6.2.2 Extraction des motifs séquentiels fréquents (MSF)

Une première modalité de réduction du nombre total de motifs, et qui correspond usuellement au désir de l'utilisateur, est d'extraire seulement les motifs qui dépassent un seuil donné de

fréquence d'occurrences. C'est le premier type de contrainte anti-monotone, la contrainte de fréquence ou de support, CS, généralement utilisée dans tous les extractions des motifs dans la littérature [10, 223, 179]. La première application d'extraction de MSF à partir des STIS a été réalisée dans [124] et [123].



a) b)

FIG. 6.4 – a) Le comportement du nombre de motifs séquentiels fréquents en fonction du seuil de support relatif, σ_{rel} et du nombre de symboles, s et b) Le comportement du temps d'extraction des motifs séquentiels fréquents en fonction du seuil de support relatif, σ_{rel} et du nombre de symboles, s.

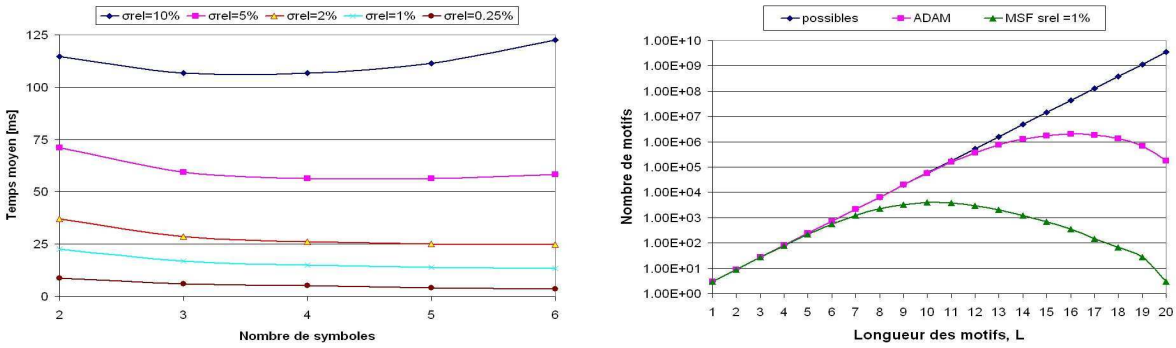
Les premiers résultats d'extraction de motifs séquentiels fréquents sont présentés dans la Figure 6.4. Le nombre de motifs croît normalement avec la diminution du seuil de support relatif et avec l'augmentation du nombre de symboles utilisés pour décrire l'évolution au niveau pixel, s (Figure 6.4a).

Dans la Figure 6.4b), le temps d'exécution présente la même dépendance, mais le taux de variation avec σ_{rel} est plus petit, fait qui explique la variation du temps moyen pour l'extraction d'un motif, présentée dans la Figure 6.5a). Le temps moyen croît avec le seuil du support relatif. Ce comportement est dû à la croissance du poids du nombre de motifs extraits (et donc écrits sur disque) par rapport au nombre de motifs visités (vérifiés) avec la diminution du seuil de support relatif. La dépendance du temps moyen d'extraction avec le nombre des symboles utilisés est décroissante avec s pour les valeurs étudiées. On observe la présence d'un minimum qui se déplace vers les petits nombres de symboles quand le seuil du support relatif croît. Les caractéristiques de la base de séquences concernant la présence réduite de motifs très fréquents font que, pour valeurs grandes de σ_{rel} et de s, le temps moyen croît.

La réduction du nombre de motifs extraits avec un seuil de support σ_{rel} en comparaison du nombre possible de motifs et du nombre de motifs existants dans la base de séquences est explicitée, en fonction de leurs longueurs, dans la Figure 6.5b). Le nombre de motifs longs se réduit considérablement et le maximum de la distribution de motifs séquentiels fréquents se déplace vers des petites longueurs.

La caractéristique de transmission de la réduction faite par l'extraction des motifs séquentiels fréquents par rapport aux motifs séquentiels existants dans la base de données (Figure 6.6a) est celle d'un filtre équivalent passe bas qui transfère des signaux en fonction de leur longueur. La pente de coupure se déplace vers les petites longueurs avec l'augmentation du seuil de support relatif et du nombre de symboles pour les valeurs des pixels. Ces courbes traduisent le fait que le pourcentage des motifs séquentiels fréquents extraits par rapport aux motifs séquentiels de la base de séquences ADAM se réduit avec la longueur et que cet effet est accéléré pour de grandes valeurs du support et du nombre de symboles.

Les résultats de la distribution des motifs suivant leurs longueurs, pour les paires de pa-

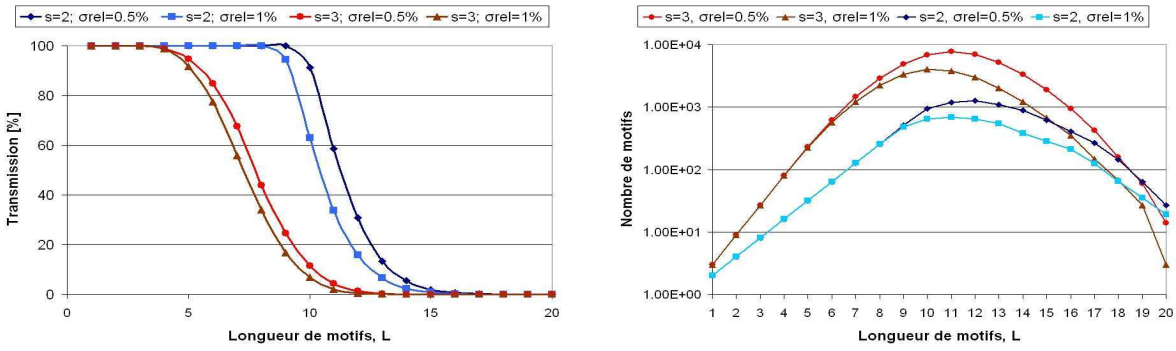


a) b)

FIG. 6.5 – a) Temps moyen d'extraction d'un MSF en fonction du nombre de symboles, s, et du seuil de support relatif, σ_{rel} et b) Les dépendances des nombres de motifs séquentiels possibles, existants dans la base de données et fréquents, en fonction de leurs longueurs ($s = 3$).

ramètres s et σ_{rel} de la Figure 6.6a), sont présentés dans la Figure 6.6b). Le déplacement et la diminution des maximums vers de petites longueurs avec la croissance du seuil de support relatif peuvent être remarqués. Avec la diminution du nombre des symboles, les maximums se déplacent vers de grandes longueurs. Cet effet conduit à l'obtention de plusieurs motifs de longueur maximale pour $s = 2$ par rapport à $s = 3$. La quantification avec $s = 2$ conduit à cet effet d'inversion.

À partir de la comparaison de ces deux derniers graphiques (Figure 6.6 a et b) il est évident que le décalage entre la pente de la caractéristique de transmission et le maximum de la distribution s'accentue avec la croissance du nombre de symboles, s.



a) b)

FIG. 6.6 – a) Les fonctions de transfert équivalent pour le processus d'extraction des MSF par rapport aux MS et b) Distributions des MSF selon leurs longueurs.

De point de vue de la mémoire utilisée, la variation de celle-ci avec le seuil de support relatif est très lente. La mémoire impliquée croît avec la diminution du nombre de symboles parce que l'algorithme utilise de bases projetées. En effet, les symboles deviennent très fréquents, et l'algorithme peut atteindre de nombreux motifs de taille 20, ce qui représente 20 bases projétées en mémoire. Ainsi, la mémoire utilisée a la valeur de 2,23 GB pour $s = 6$ et croît à 4,26 GB pour $s = 2$.

Une caractéristique de cette base de données est la présence de l'eau dans la scène d'ADAM, un objet large comprenant plus de 2% des pixels d'une image. Les pixels de l'eau ont la valeur d'IVDN minimale et un degré élevé de connexité. La conséquence est que pour $\sigma_{rel} <= 2\%$ et pour toutes les valeurs de s considérées (2 à 6) il y a au moins un motif complet de la forme

1×20, correspondant à l'eau. Pour le reste des objets la longueur des motifs diminue avec la croissance de s et de σ_{rel}. Pour des valeurs grandes de s, les pixels des autres "objets" de la scène ont des évolutions fréquentes plus courtes. Par exemple, pour $s = 6$ et $\sigma_{rel} = 2\%$ le motif le plus long après le motif de l'eau, de longueur 20, et différent de l'eau, est de longueur 12. Pour $\sigma_{rel} > 2\%$, on peut remarquer aussi l'effet de la compression dû à la réduction du nombre de symboles, s. Le lissage impliqué par le petit nombre de symboles en comprimant/regroupant les évolutions non fréquentes peut conduire à des évolutions résultantes qui cumulent les supports et dépassent le seuil de support dans ce cas. Plus le nombre de symboles est petit, plus la longueur maximale des motifs est grande.

Les comportements du nombre de motifs, de leurs longueurs et du temps d'extraction avec la variation du seuil σ_{rel} et du nombre de symboles, s, sont classiques [10, 223, 179]. Une possibilité de caractériser les dépendances du nombre de motifs est constituée par l'équivalence du processus d'extraction avec l'action d'un filtre passe bas dans le domaine des longueurs des motifs. La présence de l'eau (plus de 2% de l'image) parmi les objets de la scène de la STIS ADAM, donne des dépendances spécifiques pour la longueur maximale des motifs.

6.2.3 Extraction de motifs séquentiels fréquents groupés (MSFG) avec la contrainte sur connexité moyenne (CM)

Pour mettre en évidence des ensembles des pixels qui partagent la même évolution temporelle et qui ont un degré élevé de connexité, la notion de MSFG a été définie dans le chapitre 4 (définition 4.9). Ce type de motifs permet de trouver des régions thématiques de la scène observée et peut offrir de bons candidats pour un éventuel clustering.

La mesure anti-monotone servant de base à la contrainte sur connexité globale (définition 4.5) est similaire à la première mesure de ce type, la fréquence ou le support. Sur la base de cette mesure de connexité ont été définies les mesures de connexité moyenne, CM, et de connexité relative au support minimum, CRSM [120, 126, 119, 160, 116, 121, 122]. L'introduction de ces notions liées à l'aspect spatial des motifs va permettre une meilleure caractérisation thématique de la scène observée. L'implémentation comme contraintes actives des contraintes sur les mesures anti-monotones de connexité (CG ou CRSM) va améliorer le processus de fouille des données au niveau de la consommation de ressources.

L'objet de cette sous-section est de mettre en exergue comment l'extraction de MSFG dépend du nouveau paramètre, le seuil de la mesure de CM, κ, et des paramètres antérieurs L, σ_{rel} et s.

La distribution par longueur des motifs extraits avec la contrainte de CG, présentée dans la Figure 6.7, a l'allure usuelle de cloche avec les maximums situés dans la région de longueurs intermédiaires. Les valeurs des maximums croissent et leurs positions se déplacent vers des longueurs plus grandes avec la diminution de la connexité globale. Les valeurs élevées de la connexité globale utilisée comme paramètre de contrôle et sa relative faible signification pour l'utilisateur ont conduit à la définition de la connexité moyenne, CM (définition 4.7). Cette mesure a une signification spéciale et elle peut être utilisée dans le post-traitement de l'extraction, pour réduire le nombre de motifs séquentiels fréquents extraits.

La dépendance suivante, représentée pour la variation du nombre de MSFG, est celle suivant la longueur du motif ($N_m(L)$), en comparaison avec les autres types de motifs obtenus jusqu'à maintenant (Figure 6.8a). Pour la STIS ADAM les nombres des motifs (pour toutes les longueurs) décroissent dans l'ordre MS, MSF et MSFG. Pour s et σ_{rel} constants, la croissance du seuil de connexité moyenne, κ, conduit à la diminution attendue pas seulement du nombre total de MSFG extraits, mais aussi du nombre de motifs de chaque longueur.

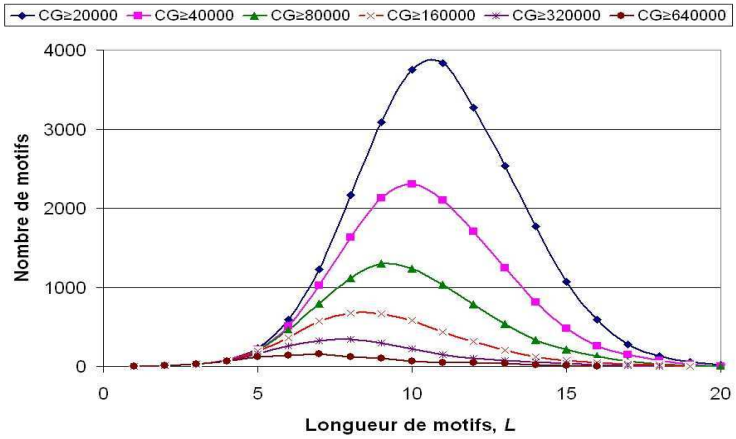
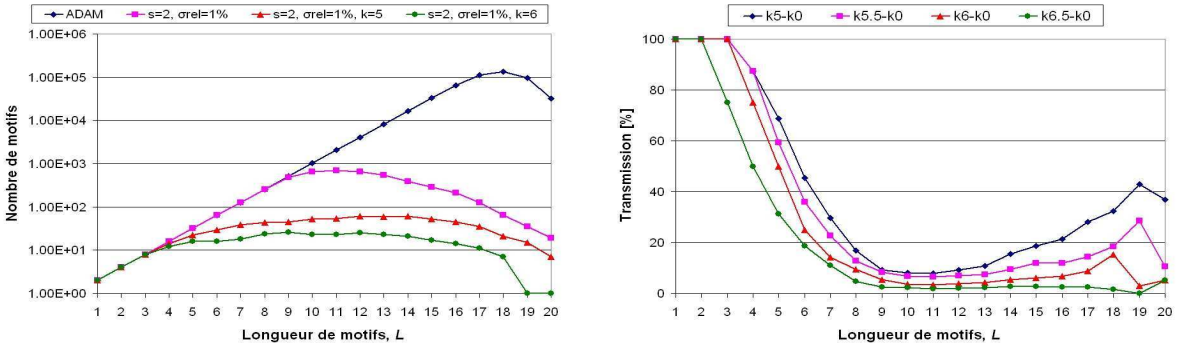


FIG. 6.7 – La distribution par longueur des MSF suivant leur connexité globale, CG ($s = 3$, $\sigma_{rel} = 1\%$).



a) b)

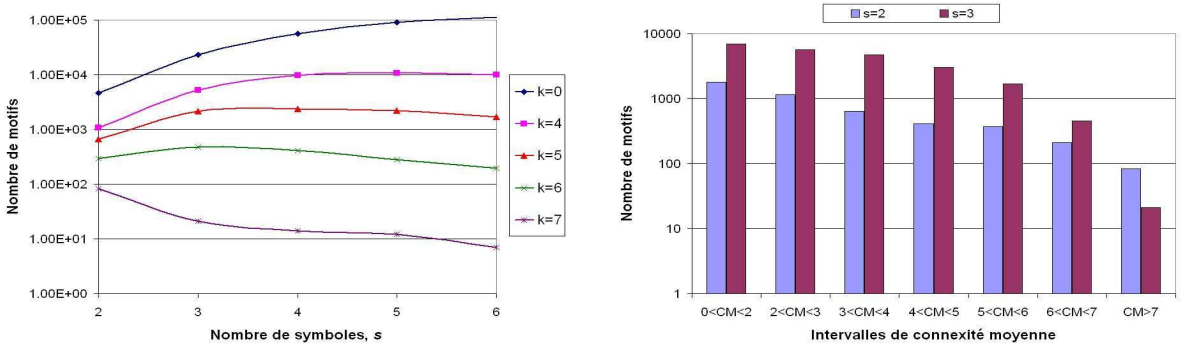
FIG. 6.8 – a) Comparaison des distributions suivant la longueur des motifs de la BS - ADAM, MSF et MSFG ($s = 2$, $\kappa = 5$ et $\kappa = 6$ pour $\sigma_{rel} = 1\%$) et b) Les caractéristiques de transmission entre les MSF et les MSFG ($s = 2$, $\sigma_{rel} = 1\%$).

En général, le nombre de MSFG extraits diminue avec l'augmentation du seuil de support relatif σ_{rel} et du seuil de connexité moyenne, κ. L'influence de la variation du seuil σ_{rel} sur le nombre de MSFG est très faible pour des valeurs moyennes et grandes de κ en démontrant qu'il y a de formations thématiques terrestres compactes et vastes qui se retrouvent pour n'importe quel $\sigma_{rel} \leq 2\%$.

Les caractéristiques de transmission pour un domaine plus réduit de valeurs de la CM (5-6,5) mais intéressant pour des applications présentent une zone «fenêtre», qui permet le passage spécialement des motifs longs, passage plus accentué pour des κ petits.

La Figure 6.8b) permet l'observation du comportement des motifs longs qui sont d'intérêt pour caractériser des évolutions. La fonction de transmission décrit un comportement d'un filtre passe bas qui élimine les motifs en commençant des longueurs d'autant plus courtes que le seuil κ croît. Une situation favorable pour les motifs longs grâce à leur degré élevé de connexité peut être remarquée. Ce fait est la conséquence de l'organisation spécifique à la thématique de la scène observée et du degré de connexité des motifs longs.

La variation du nombre de MSFG a un comportement particulier en fonction de la valeur du seuil de connexité moyenne, κ, et du nombre des symboles, s, comportement mis en exergue par la Figure 6.9a). Pour une extraction de type antérieur, qui ne tient pas compte de la connexité

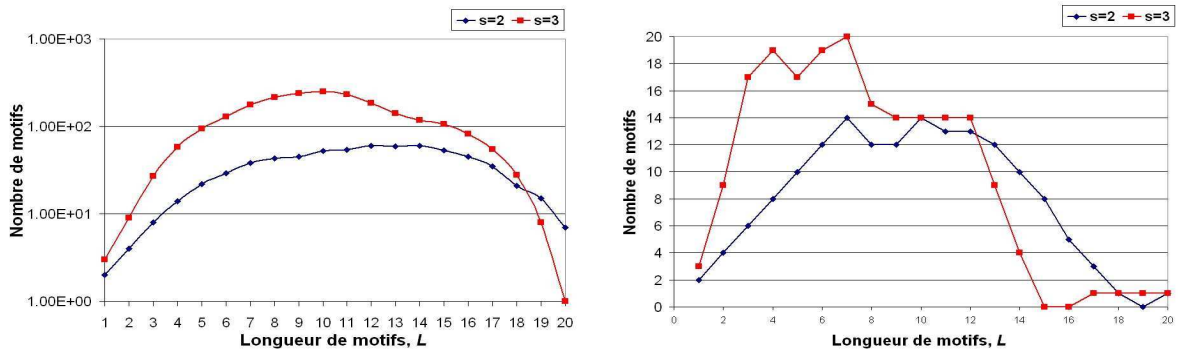


a) b)

FIG. 6.9 – a) La dépendance du nombre de MSFG suivant la discrétisation, s, et le seuil de connexité moyenne, κ ($\sigma_{rel} = 1\%$) et b) La répartition des MSFG suivant leur connexité moyenne ($\sigma_{rel} = 1\%$, $s = 2$ et $s = 3$).

moyenne, (la courbe $\kappa = 0$), le nombre de motifs croît avec le nombre de symboles, s. Pour des valeurs petites et moyennes du seuil κ de la CM, la dépendance présente un maximum qui se déplace vers un nombre réduit de symboles pour l'augmentation du seuil κ. Par exemple, pour $\kappa = 5$ et $\kappa = 6$, valeurs d'intérêt dans notre étude, le nombre maximum de motifs est offert par la discrétisation $s = 3$, la description la plus usuelle et compréhensible pour l'utilisateur (des valeurs de pixels petites, moyennes et grandes). Pour une connexité très élevée, $\kappa = 7$, le nombre de motifs diminue avec l'augmentation de s, fait qui est en concordance avec les cas d'inversions montrés plus loin (par exemple dans la Figure 6.9b).

Si, en général, un nombre plus grand de symboles utilisés, s, assure un nombre plus grand des MSFG, la Figure 6.9b) montre que pour $6 < CM < 7$ l'extraction donne presque le même résultat que pour $s = 2$ et $s = 3$. Pour $CM \geq 7$ a lieu une inversion du comportement, la quantification plus forte de $s = 2$ donnant plus de motifs très connexes.



a) b)

FIG. 6.10 – a) La distribution par longueur des MSFG pour $s = 2$ et $s = 3$ ($\sigma_{rel} = 1\%$, $\kappa = 5$) et b) La distribution du nombre de MSFG suivant leur longueur pour $s = 2$ et $s = 3$ ($\sigma_{rel} = 1\%$, $\kappa = 6,5$).

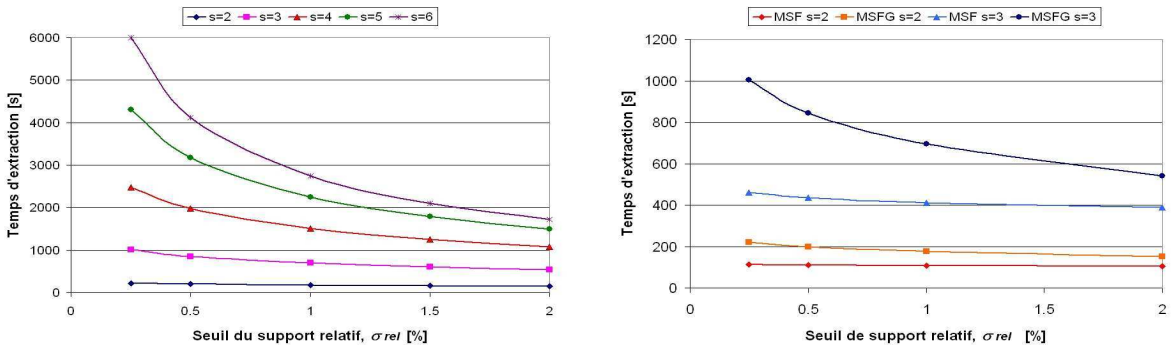
Une situation similaire d'inversion est présentée dans la Figure 6.10a), où la quantification plus accentuée de $s = 2$ assure plus de motifs connexes de longueur 19 et 20 que $s = 3$. La binarisation des valeurs des pixels assure une distribution plus aplatie et d'intérêt pour l'utilisateur : un nombre total de motifs réduit mais le plus grand nombre de motifs de longueur maximale ($L = 20$).

Pour des degrés élevés de la connexité moyenne où l'inversion est plus accentuée, des si-

tuations dans lesquelles on voit qu'une contrainte de seuil minimum sur cette mesure n'est pas anti-monotone peuvent être obtenues (Figure 6.10b). La courbe rouge ($s = 3$) montre en effet qu'il n'y a pas des motifs pour $L = 15$ et $L = 16$, alors que des motifs sont présents pour $L \geq 17$. En outre, l'inversion entre les nombres de motifs pour $s = 2$ et $s = 3$ commence plus tôt, à la longueur 13.

Concernant la variation de la longueur maximale avec les paramètres de contrôle σ_{rel}, s et κ, les expérimentations démontrent l'indépendance de cette longueur suivant la variation du seuil σ_{rel} (dans le domaine étudié 0.25%-2%), fait qui prouve l'existence d'objets dans la scène ayant une surface qui dépasse la gamme considérée du seuil de support (voir aussi les résultats sur les MSF, sous-section 6.2.1). Pour un κ et s donnés, la longueur maximale est la même. Il y a des comportements différents entre la situation correspondante à des petites et moyennes valeurs du seuil de la CM et la situation de grandes valeurs. Pour $\kappa \leq 6,5$ et pour toutes les valeurs de s la longueur maximale est de 20. Pour la valeur $\kappa = 7$ et pour toutes les valeurs expérimentées de σ_{rel}, la longueur maximale commence à diminuer jusqu'à la valeur de 4 avec la croissance de s (dans la gamme 2-4). Pour des valeurs grandes de $s \geq 5$ la longueur redevient grande ($L = 18$). Le comportement anormal pour de valeurs grandes de s est provoqué par la présence de l'eau dans la scène. L'eau a des valeurs spectrales dans Rouge et PIR qui conduisent à un IVDN presque nul. Pour des s petits, la quantification apporte d'autres pixels aussi dans la classe avec la valeur minimale d'IVDN, pixels différents de l'eau, qui ne sont pas très connexes probablement. Dans cette situation, il est difficile d'obtenir des motifs longs qui ont aussi un grand degré de connexité. Quand le nombre de symboles croît, dans la classe avec des valeurs minimales d'IVDN, le poids des pixels correspondants à l'eau croît et on peut atteindre des grands degrés de connexité et longueurs (fait caractéristique pour l'eau). Quand le degré de connexité est environ la valeur 7,5 le comportement revient à la normale parce que la connexité moyenne du l'eau de la rivière est dépassée ; la longueur maximale décroît drastiquement et dépend du nombre de symboles. Ainsi, la longueur maximale décroît de 6 à 1 pour la variation de s entre les valeurs 2 et 6.

Ces explications sont en accord avec les valeurs de la connexité moyenne pour les motifs longs de valeur «1» qui correspondent aux pixels couverts par l'eau. Pour $s = 2$ la $CM = 6,64$ et elle croît continuellement, $CM = 6,95$ pour $s = 3$, $CM = 6,99$ pour $s = 4$. Pour les valeurs 5 et 6 du nombre de symboles, la CM est située dans la région 7-7,5.



a) b)

FIG. 6.11 – a) Temps d'extraction des MSFG en fonction de σ_{rel} et s et b) La comparaison entre les temps d'extraction des MSF et MSFG.

Les temps d'extraction des MSFG ne dépendent pas de la valeur du seuil de connexité moyenne, κ, (Figure 6.11a), cette opération étant seulement un post-traitement. De cette manière les courbes sont valables pour toutes les valeurs de κ. Les MSF extraits sont filtrés avec la condi-

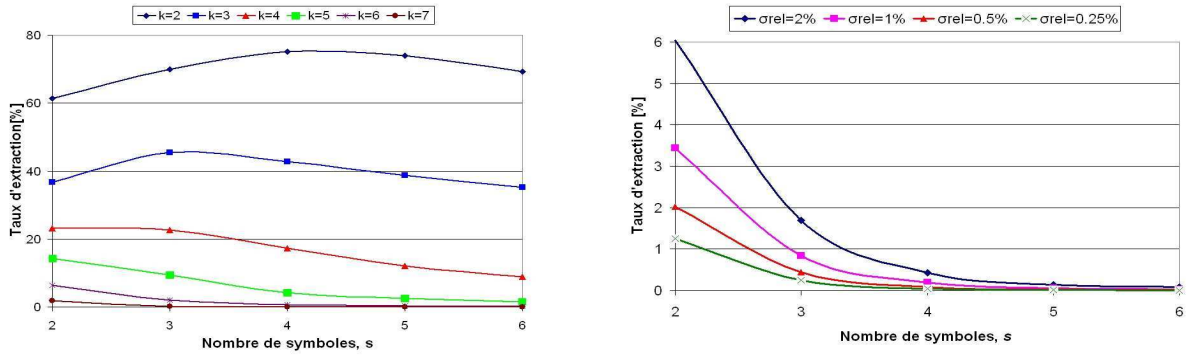
tion que leur degré de connexité dépasse un seuil. Ces temps décroissent avec l'augmentation du seuil σ_{rel} et avec la diminution du nombre de symboles, s. À cause des calculs impliqués par ce post-traitement, les temps d'exécution sont plus grands que ceux pour les MSF. Une comparaison entre les temps pour les extractions des MSF et MSFG et présentée dans la Figure 6.11b). On peut définir un facteur d'amplification du temps d'extraction (t_{MSFG}/t_{MSF}) qui décroît avec l'augmentation du seuil de fréquence σ_{rel} et avec la diminution du nombre de symboles, s.

Il est utile d'introduire le niveau de restriction de la contrainte appliquée comme le rapport entre les nombres de MSFG et MSF.

Définition 6.1. ***(taux d'extraction)*** Le taux d'extraction (ou simplement l'extraction) est le rapport entre le nombre de motifs extraits et le nombre de motifs visités pour vérifier la contrainte.

Avec nos notations, ce taux d'extraction est défini par le rapport N_m/N_{vis}, où N_m est le nombre de motifs extraits et N_{vis} est le nombre de motifs visités pour vérifier l'accomplissement de la contrainte, et il mesure, pour un type d'extraction donné, l'efficacité de l'extraction. Les valeurs de ce taux sont comprises entre 0 et 1. Une valeur petite du taux signifie une extraction efficace.

Pour un seuil de support donné, ce taux d'extraction présente (Figure 6.12a), en fonction de s, une dépendance avec des maximums qui se déplacent vers des s grands pour des κ réduits et une décroissance continue pour κ grands, usuellement le domaine d'intérêt.



a) b)

FIG. 6.12 – a) La variation du taux d'extraction avec s et κ (σ_{rel}=1%) pour l'extraction de MSFG pour $\sigma_{rel} = 1\%$ et b) La dépendance du taux d'extraction suivant le nombre de symboles, s, et le seuil de support σ_{rel} pour le seuil de CM, $\kappa = 6,5$.

Pour des valeurs grandes du seuil de connexité ($\kappa = 6.5$), la Figure 6.12a) montre la croissance du taux d'extraction avec la diminution de s et l'augmentation de σ_{rel}. Pour des valeurs réduites du nombre de symboles s, plusieurs MSFG sont préservés. Cette afirmation est valable pour le nombre total de motifs comprenant toutes les longueurs.

L'introduction de la mesure de CM et le post-traitement impliqué par la contrainte correspondante, permet une réduction supplémentaire du nombre de motifs séquentiels extraits. Ces motifs ont une signification spatiale, les pixels couverts ayant un attribut de connexité, propriété qui assure un rôle de très bons candidats si un regroupement du contenu thématique de la scène est désiré. Le seul désavantage est constitué par l'implémentation passive de la contrainte, le fait que son introduction conduit à un post-traitement consommateur de temps de calcul.

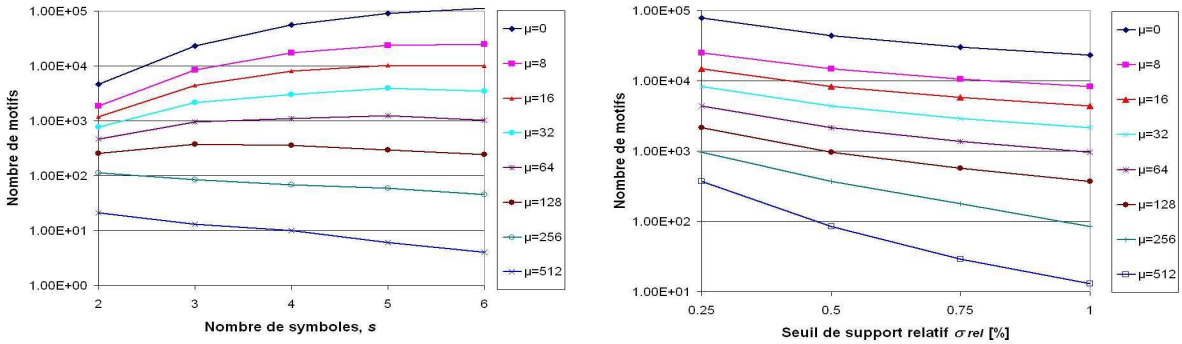
6.2.4 Extraction avec la contrainte sur connexité relative au support minimum (CRSM)

L'alternative avec implémentation de la connexité relative au support minimum, CRSM (définition 4.10), surpasse le désavantage d'une implémentation passive. CRSM étant une mesure anti-monotone, son implémentation active dans la fouille est possible.

Le principal avantage de l'extraction avec la contrainte active concernant la connexité relative au support minimum, CRSM, est la réduction du temps d'extraction assurée par la réduction du nombre de motifs visités, conséquence de l'élagage efficient.

Le nombre de ces motifs se réduit en comparaison avec celui de MSF à cause de l'implémentation dans le processus d'extraction de la contrainte active $CRSM > \mu$, où μ est le seuil pour ce type de connexité. Le comportement graphique de la dépendance du nombre de motifs suivant les paramètres usuels s, σ_{rel} et μ est le même que celui pour l'extraction qui utilise la connexité globale, CG, comme support de contrainte.

La dépendance décrite par la Figure 6.13a) est semblable à celle qui montre l'influence du seuil κ sur le nombre de motifs (Figure 6.9). Pour des valeurs réduites et moyennes de la CRSM, le nombre de motifs a le comportement normal de croissance avec le nombre de symboles s. Les maximums des courbes décroissent en valeur et se déplacent vers des s petits pour la croissance du seuil du degré de connexité, μ (pour des valeurs usuelles de μ). Pour un μ très grand (256 et 512 pour la Figure 6.13a) le nombre de motifs décroît avec l'augmentation de s. C'est un fait qui tient de la particularité de la statistique de données de la STIS ADAM. Le nombre de motifs de connexité extrême est très réduit et la même chose peut être affirmée sur leur longueur.



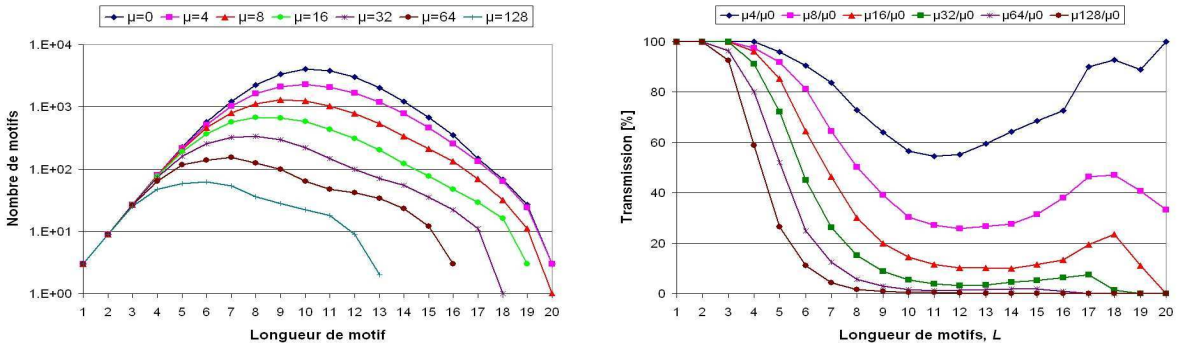
a) b)

FIG. 6.13 – a) La dépendance du nombre de motifs suivant s et μ ($\sigma_{rel} = 1\%$) et b) La dépendance du nombre de motifs suivant σ_{rel} et μ ($s = 3$).

Concernant la dépendance du nombre de MSF extraits avec la contrainte sur CRSM suivant le seuil de fréquence, (Figure 6.13b), on obtient des comportements normaux, ce nombre décroît avec l'augmentation de σ_{rel} et de μ. Plus fortes sont ces conditions, plus petit est le nombre de motifs.

Pour les buts de recherche énoncés antérieurement, il est nécessaire d'étudier non seulement le nombre de motifs mais aussi leur distribution par longueur. La Figure 6.14a) présente cette distribution pour deux valeurs usuelles des paramètres s et σ_{rel}. Ici, sont observés clairement la diminution du nombre de motifs pour chaque longueur et le déplacement du maximum de la distribution vers les longueurs petites avec l'augmentation du seuil de CRSM. Ainsi, on peut s'attendre que le nombre des motifs longs diminue avec l'augmentation du seuil μ de CRSM.

La bonne connexité de MSF extraits est prouvée par la superposition des courbes pour $\mu = 0$



a) b)

FIG. 6.14 – a) La distribution des motifs selon leur longueur, L, et leur seuil de CRSM, μ ($s = 3$ et $\sigma_{rel} = 1\%$) et b) Les caractéristiques de transmission des MSF par rapport aux motifs extraits avec la contrainte de CRSM suivant la longueur et le seuil de CRSM ($s = 3$ et $\sigma_{rel} = 1\%$).

et $\mu = 4$ pour les longueurs les plus élevées ($L \geq 18$).

En faisant la comparaison entre les motifs extraits avec la contrainte sur CRSM et les MSF, on peut tracer les caractéristiques de transmission pour le passage de l'extraction avec un seule contrainte anti-monotone basée sur la fréquence à l'extraction qui englobe aussi la deuxième contrainte anti-monotone, celle basée sur la connexité relative au support minimum (Figure 6.14b).

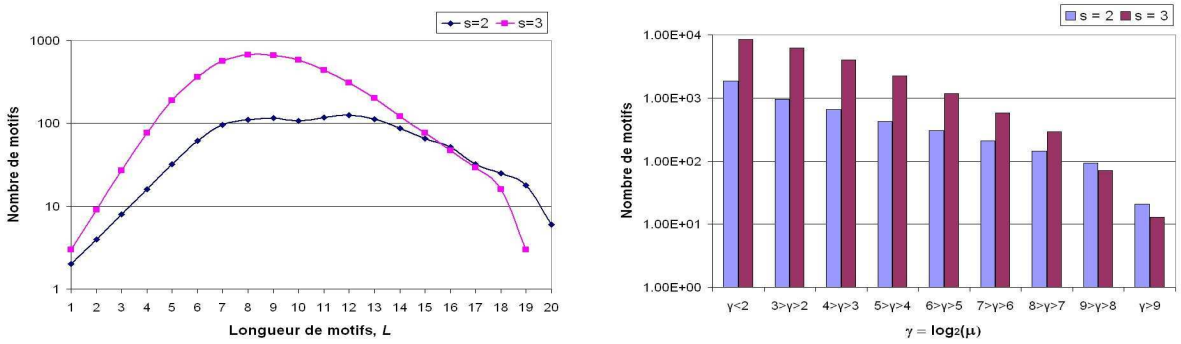
Pour des valeurs très grandes du seuil de CRSM, μ, la caractéristique de transmission est très nette, semblable à un filtre passe-bas, et commence à couper à des longueurs d'autant plus courtes que μ croît. Avec la diminution du seuil μ de CRSM, la caractéristique coupe à des longueurs plus grandes et son allure se déforme, en permettant à plus des motifs longs de passer. Pour des seuils μ très petits, la caractéristique devient semblable à un filtre coupe bande pour les longueurs intermédiaires.

Le comportement de «fenêtre» pour des longueurs grandes, semblable à celui présenté dans la Figure 6.8 pour l'influence du seuil κ, est présenté. Les valeurs élevées de la caractéristique de transmission pour $\mu = 4$ et $L \geq 17$ montre le fait mis en évidence également par la Figure 6.14 : pour les longueurs grandes, les MSF ont la valeur de la CRSM d'environ 4 (la petite différence entre les courbes correspondantes aux $\mu = 0$ et $\mu = 4$).

Si la longueur des motifs extraits est étudiée, des résultats d'inversions sont obtenus, comme ceux présentés dans la Figure 6.15a). La quantification extrême obtenue pour $s = 2$ conduit à un nombre supérieur de motifs très longs en comparaison avec le nombre obtenu pour le cas $s = 3$.

Le même choix de la valeur $s = 2$ assure les meilleurs résultats même de point de vue du degré de connexité (voir la Figure 6.15b). Donc, il est nécessaire d'avoir une quantification extrême si nous voulons obtenir des motifs longs avec un degré de connexité élevé.

La longueur maximale des motifs est 20 pour le seuil de connexité $\mu < 8$ et pour toutes les valeurs considérées de s $(2 - 6)$ et σ_{rel} $(0 - 2\%)$ dans les domaines étudiés. La valeur $\mu = 8$ est atteinte, pour le seuil $\sigma_{rel} = 2\%$ seulement pour $s = 2$, la quantification la plus favorable. Les pixels couverts par l'eau ont un support d'environ 2% pour une longueur de 20, et un degré de connexité moyenne d'environ 7 et, parce que $CRSM = CM \times supp/\sigma$, leur connexité relative au support minimum est moindre que la valeur 8 pour un seuil $\sigma_{rel} = 2\%$ et pour $s > 2$. Le résultat est que la longueur décroît jusqu'à une valeur pour laquelle le support atteint une valeur suffisante pour obtenir $CRSM \geq 8$. Ainsi, les valeurs de L_{max} sont 19 pour $s = 3$ et 18 pour s dans la gamme $4 - 6$. Pour des valeurs du seuil μ supérieures à 8, le comportement normal est

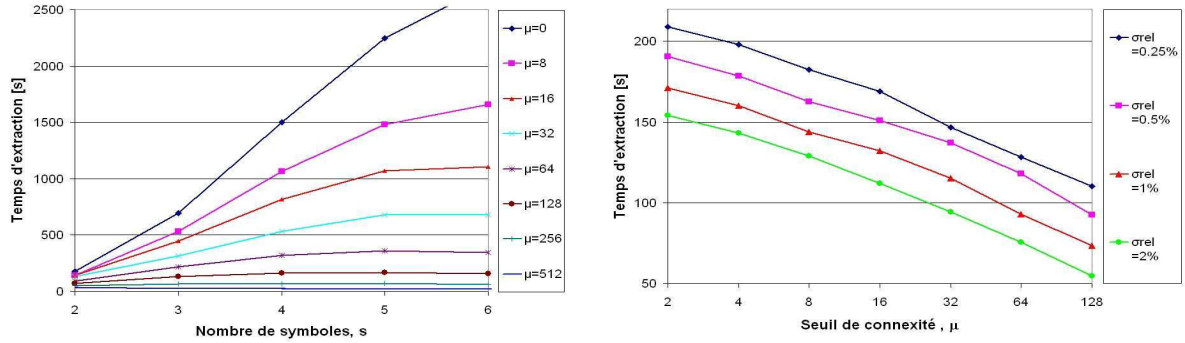


a) b)

FIG. 6.15 – a) La distribution de motifs selon leur longueur pour $s = 2$ et $s = 3$ ($\sigma_{rel} = 1\%$ et $\mu = 16$) et b) La distribution des motifs selon leur degré de connexité pour $s = 2$ et $s = 3$ ($\sigma_{rel} = 1\%$).

rétabli : la longueur maximale décroît avec l'augmentation de μ, s et σ_{rel}.

Les temps d'extraction obtenus avec l'implémentation active de la contrainte liée à la CRSM sont plus courts en comparaison avec l'extraction filtrée avec la contrainte correspondante à la connexité moyenne, CM. Le temps d'extraction croît avec l'augmentation du nombre de symboles, s, et la diminution du seuil de CRSM, μ (Figure 6.16a) et du seuil de fréquence, σ_{rel} (Figure 6.16b).



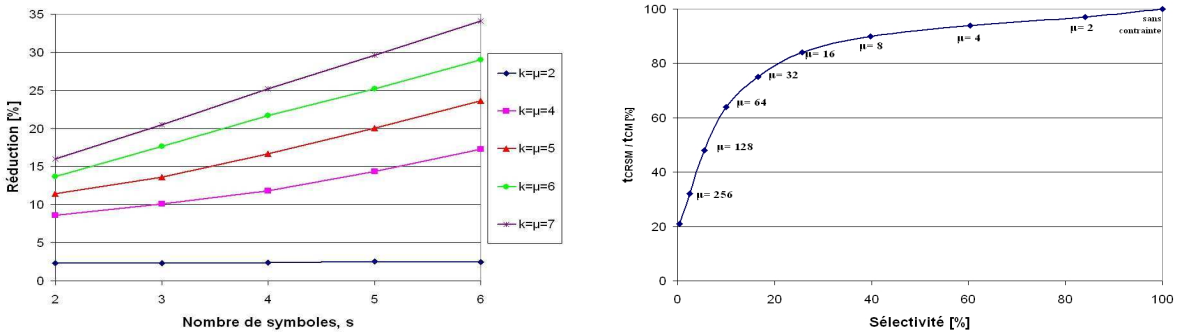
a) b)

FIG. 6.16 – a) La dépendance du temps d'extraction suivant s et μ ($\sigma_{rel} = 1\%$) et b) La dépendance du temps d'extraction suivant μ et σ_{rel} ($s = 2$).

Des valeurs raisonnables pour les paramètres envisagés de calcul sont remarquées : un nombre réduit de symboles, s, (pour la quantification réduite des descriptions des évolutions des pixels), un seuil relatif de fréquence dans la zone de 1% et un seuil de connexité élevé.

La principale conséquence de l'implémentation de la contrainte active basée sur la connexité relative au support minimum est la réduction du temps d'extraction $(1 - \frac{t_{CRSM}}{t_{CM}})$ en comparaison avec le cas de la connexité moyenne avec les mêmes seuils μ et κ (Figure 6.17a).

Si les seuils des extractions basées sur la connexité moyenne, CM, et sur la connexité relative au support minimum, CRSM, sont égalisés, une comparaison entre les temps d'extraction et un calcul de la réduction de ce temps dans le cas d'utilisation de la condition de la contrainte anti-monotone, $CRSM > \mu$, peuvent être faits. Dans la Figure 6.17a), on peut voir que cette réduction calculée en pourcents croît avec l'augmentation du nombre de symboles, s, et les seuils $\kappa = \mu$. En effet, la réduction peut être plus grande si des valeurs plus grandes pour le seuil,



a) b)

FIG. 6.17 – a) La réduction du temps d'extraction entre CRSM et CM ($\sigma_{rel} = 1\%$) et b) La réduction du temps d'extraction CRSM vs CM suivant la sélectivité pour $s = 2$ et $\sigma_{rel} = 1\%$.

μ, de la CRSM sont utilisées. C'est le plus grand problème de cette mesure de connexité : sa signification est un peu confuse pour utilisateur.

Afin de mieux comprendre le comportement d'une extraction (2) avec une contrainte plus restrictive, C_2, en faisant la comparaison avec une extraction (1) qui a un autre type de contrainte C_1, les suivantes définitions sont introduites. Les nouvelles notions permettent d'exprimer d'une manière plus concise et suggestive les dépendances des paramètres étudiés.

Définition 6.2. ***(taux de sélectivité)*** Le taux de sélectivité (ou simplement la sélectivité) est le rapport entre le nombre de motifs satisfaisant la contrainte $C_1 \wedge C_2$ et le nombre de motifs satisfaisant seulement la contrainte C_1.

Cette définition 6.2 tient compte seulement des motifs satisfaisant C_1 plutôt que de tous les motifs présents dans la base de séquences ADAM afin de mieux analyser l'efficacité de l'extraction avec la contrainte C_2. De cette manière, la mesure reflète la proportion de motifs qui satisfont séparément les deux contraintes et elle est comprise entre 0 et 1. Plus la sélectivité est proche de 1, moins la contrainte C_1 est sélective car tous les motifs satisfaisant C_2 satisfont également C_1. À l'inverse, lorsque la sélectivité est proche de 0, la contrainte C_1 est plus sélective car peu de motifs sont extraits parmi ceux satisfaisant C_2.

Dans cette section est réalisée une comparaison entre les motifs extraits avec la contrainte anti-monotone sur CRSM et ceux extraits avec la contrainte de support qui vise seulement le support ou ceux extraits avec des contraintes de support et sur CM.

Définition 6.3. ***(taux de succès de l'élagage)*** Le taux de succès de l'élagage (ou simplement l'élagage) est le rapport entre le nombre de réussites de l'élagage et le nombre des tentatives.

Le taux de succès de l'élagage (ou simplement l'élagage) est donc compris entre 0 et 1 et rend compte de l'efficacité de l'élagage. Plus le taux est grand, plus l'élagage est efficace. Dans cette section concernant la contrainte basée sur CRSM, ce taux est défini comme :

L'élagage $= 1 - N_m/N_{vis} = 1-$ taux d'extraction

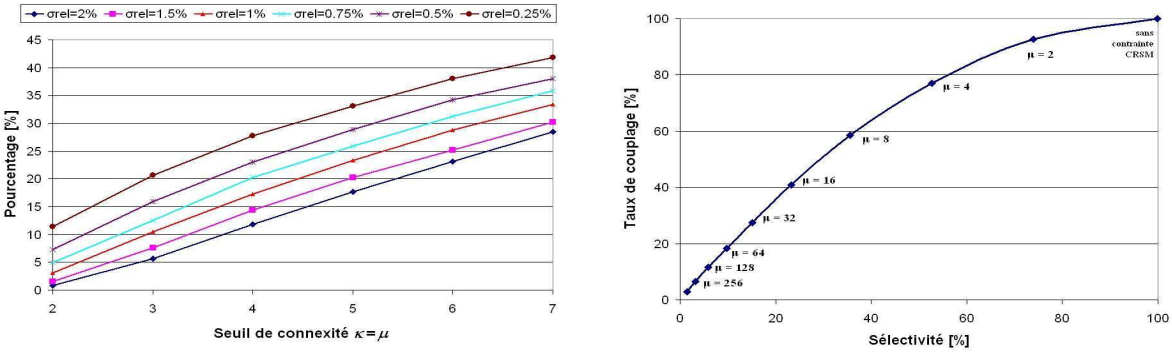
où N_m et N_{vis} ont les significations énoncées précédemment.

Définition 6.4. ***(taux de couplage du traitement)*** Pour une succession d'opérations enchaînées pour extraction, le taux de couplage de traitement (ou simplement le couplage) est le rapport $(\frac{N_{vis2}}{N_{m1}})$ entre le nombre de motifs visités pour une opération 2 (N_{vis2}) et le nombre de motifs extraits offerts par la précédente opération 1 (N_{m1}).

Les valeurs de ce taux sont comprises entre 0 et 1. Plus les valeurs se rapprochent de 1, plus le couplage est fort et il y a beaucoup de motifs à traiter. Pour un post-traitement (le cas de l'extraction avec la contrainte sur CM), le taux de couplage du traitement est de 100%.

La Figure 6.17b) illustre la plus importante propriété de l'utilisation de la contrainte sur CRSM, la réduction du temps d'extraction. Le graphique présente d'une manière concise l'influence de la croissance du seuil μ sur l'efficacité de l'extraction. On voit que pour des valeurs petites de μ la sélectivité est faible, c'est-à-dire proche de 100%.

On peut remarquer que, pour des mesures de connexité très élevées qui correspondent à une sélectivité forte, par exemple $CRSM \approx 500$, une augmentation de 5 fois de la vitesse de l'extraction peut être obtenu. Concernant la relation entre les nombres de motifs visités des ces deux types d'extraction (avec contrainte sur CRSM et sur CM), la Figure 6.18a) illustre l'augmentation de la réduction de N_{vis} avec l'agrandissement du seuil de connexité μ et la diminution du seuil de support relatif σ_{rel}.



a) b)

FIG. 6.18 – a) La réduction du nombre de motifs visités dans une extraction avec la contrainte sur CRSM vs CM suivant le seuil de connexité $\kappa = \mu$ et σ_{rel} pour $s = 2$ et b) Le taux de couplage $CRSM/CM$ vs la sélectivité pour le cas $s = 2$ et $\sigma_{rel} = 1\%$.

La réduction du temps d'extraction dépend très fortement du nombre de motifs visités par l'intermediaire du taux de couplage (Figure 6.18b). Une sélectivité forte implique un couplage réduit et, implicitement, un temps d'extraction très court.

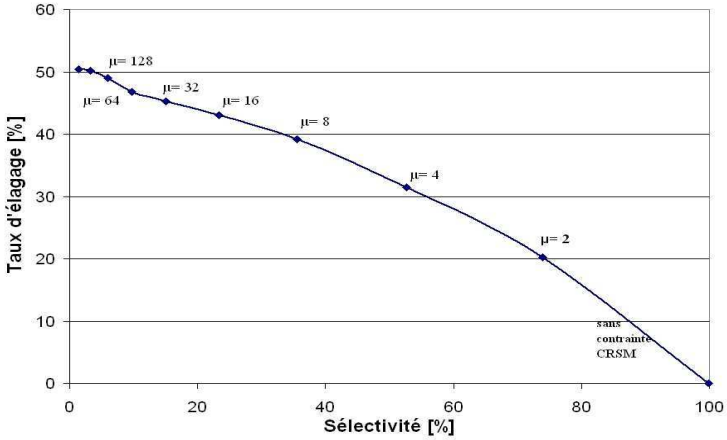


FIG. 6.19 – Le taux de succès d'élagage $CRSM/CM$ vs la sélectivité pour l'extraction avec la contrainte sur CRSM pour le cas $s = 2$ et $\sigma_{rel} = 1\%$.

Un seuil élevé de la CRSM, qui implique une sélectivité forte, peut conduire à un élagage

supérieur à 50% (Figure 6.19).

6.2.5 Extraction avec la relaxation de la contrainte sur CM par la contrainte sur CRSM ($\mu = \kappa$)

L'extraction assurée par la contrainte sur connexité moyenne (CM qui n'est pas anti-monotone) donne de bons résultats de point de vue du nombre de motifs extraits mais, étant implementée seulement par un filtrage, elle souffre des temps d'extraction plus longs. Ce type de contrainte est bien compris et peut être interprété par l'utilisateur. La contrainte basée sur la mesure de CRSM est anti-monotone et, étant implémenté activement dans le processus d'extraction, permet un élagage efficient. En conséquence, on obtient des réductions du nombre de motifs visités et du temps d'extraction. Les désavantages consistent en une insuffisante réduction du nombre de motifs extraits (pour $\kappa = \mu$) et en une signification un peu confuse de cette mesure pour l'utilisateur.

La solution consiste à tirer profit des spécificités des ces mesures en faisant des combinaisons entre elles, de type relaxation de contrainte ou conjonction. Ainsi, on peut obtenir :
- une signification assez claire pour utilisateur ;
- une bonne réduction du nombre de motifs visités par un élagage efficient ;
- une importante réduction du nombre de motifs extraits ;
- finalement, une bonne réduction du temps d'extraction.

Le premier type d'extraction, en utilisant la combinaison de ces contraintes de connexité discutées au-dessus, est constitué par l'implémentation active de la contrainte sur CRSM (avec le seuil μ), dans le processus d'extraction, suivie par le filtrage assuré par la contrainte sur CM (avec le seuil κ). Dans ce type d'extraction avec $\kappa = \mu$, l'approche accomplit la condition de complétude et de justesse par rapport à la contrainte sur CM. La complétude assure que tous les motifs de la base de données satisfaisant la condition de cette contrainte sont extraits. De cette manière, aucune information jugée pertinente pour l'utilisateur (i.e., satisfaisant sa contrainte) n'est omise. La justesse garantit que chacun des motifs extraits satisfait la contrainte sur CM.

Les résultats de l'extraction CRSM + CM impliquent que ce processus est correct et complet par rapport à la contrainte sur CM (voir la section 5.4). Dans cette section, les extractions sont réalisées avec la condition $\mu = \kappa$, le cas de la relaxation optimale.

Pour mettre en évidence les avantages de cette relaxation, des comparaisons de ce processus d'extraction sont faites avec les extractions individuelles s'appuyant sur CM et CRSM et avec l'extraction basée seulement sur la contrainte de support, contrainte de support (ou de fréquence) (CS).

Ainsi, les motifs et le nombre des motifs extraits avec la conjonction CRSM + CM ($\kappa = \mu$) sont les mêmes que ceux obtenus dans le cas d'application de la contrainte sur CM, la contrainte plus restrictive (voir le schéma de la Figure 6.20 [118]).

Les dépendances du nombre de motifs extraits par la relaxation CRSM + CM suivant les variations des paramètres s, σ_{rel} et $\kappa = \mu$, sont ainsi similaires avec les résultats présentés dans les Figures 6.9 - 6.10.

De point de vue du taux de la sélectivité qui exprime le poids des MSFG extraits parmi les motifs testés (résultant de l'application seulement de la contrainte liée au support, CS), la Figure 6.21a) présente la décroissance de cette grandeur (signifiant en effet la renforcement de la sélectivité) avec l'augmentation du degré de connexité, chose attendue parce que N_m décroît fortement avec l'augmentation du seuil de connexité. Les dépendances suivant la variation du nombre de symboles utilisés montrent l'inversion spécifique aux degrés élevés de connexité c'est-

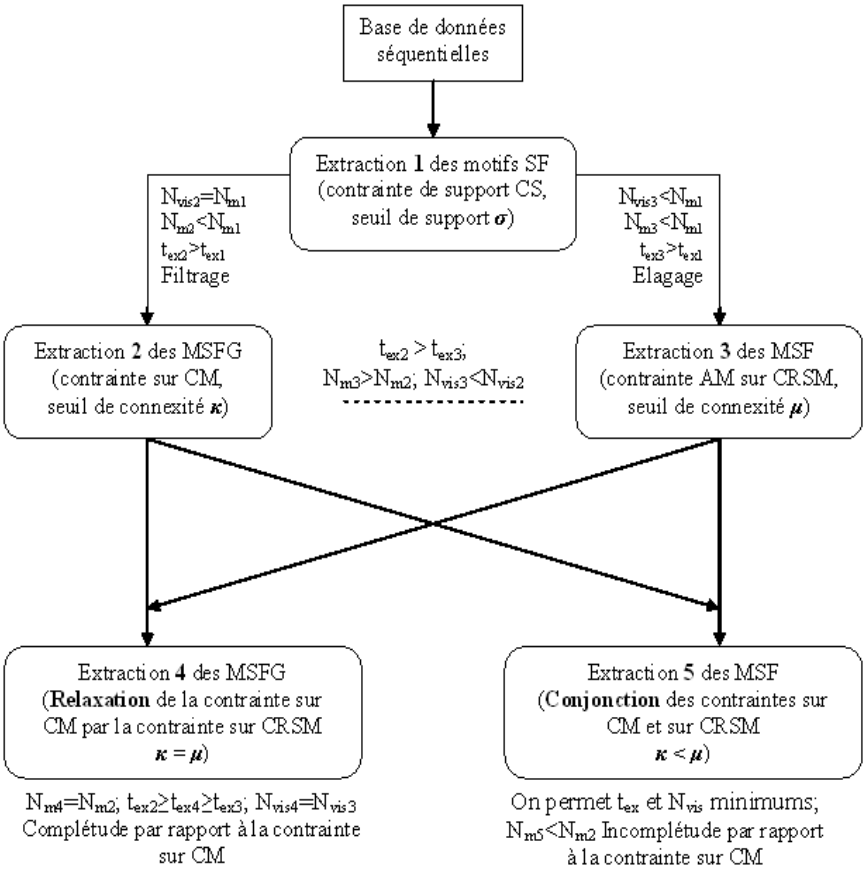
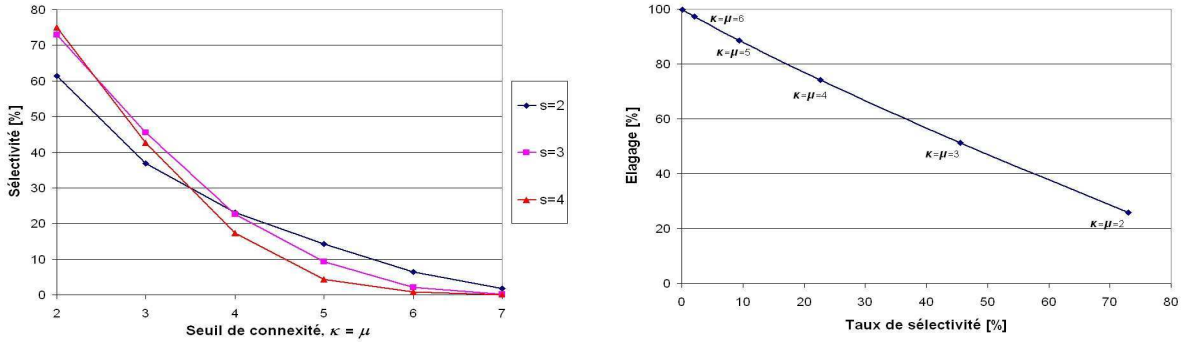


FIG. 6.20 – Schéma d'évolution des processus d'extraction de motifs

à-dire les valeurs petites de s donnent plus de MSFG. Si pour $\kappa = \mu = 2$ l'ordre de la sélectivité est $s = 2$ (la plus restrictive condition), $s = 3$, $s = 4$, pour $\kappa = \mu > 4$ l'ordre s'inverse $s = 4$, $s = 3$ et $s = 2$.

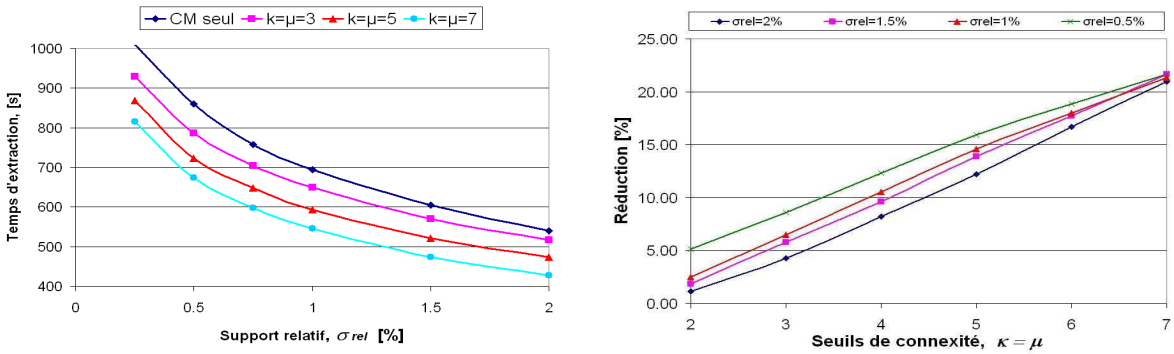


a) b)

FIG. 6.21 – a) La variation du taux de sélectivité suivant le seuil de connexité $\kappa = \mu$ et du nombre de symboles, s, pour le seuil de support $\sigma_{rel} = 1\%$ et b) La dépendance du taux de succès de l'élagage de l'extraction CRSM+CM ($\kappa = \mu$) suivant la variation du taux de sélectivité pour $\sigma_{rel} = 1\%$ et $s = 3$.

Comme une suite des fortes sélectivités correspondantes aux grandes valeurs du seuil de connexité, le taux de succès de l'élagage est important pour ces valeurs. La Figure 6.21b) met en exergue une dépendance presque linéaire entre la sélectivité et l'élagage spécifiques à l'extraction CRSM+CM. La sélectivité est d'autant plus forte que l'élagage est accentué.

Les principaux avantages de l'extraction CRSM+CM en comparaison avec celle pour CM sont les diminutions du temps d'extraction et du nombre de motifs testés, N_{vis}. La Figure 6.22a) présente la comparaison entre les temps de ces extractions et leur comportement attendu de diminution avec la croissance des seuils de support et de connexité.



a) b)

FIG. 6.22 – a) Les temps d'extraction des motifs CRSM+CM et CM suivant la variation des seuils de connexité, $\kappa = \mu$, et de support relatif, σ_{rel}, dans le cas $s = 3$ et b) La réduction du temps d'extraction en utilisant CRSM+CM ($\kappa = \mu$) suivant la variation des seuils de connexité et de support dans le cas $s = 3$.

La Figure 6.22b) mesure le gain de temps d'extraction apporté par la relaxation anti-monotone en comparaison avec l'extraction seulement avec la contrainte sur CM et illustre l'augmentation de la réduction du temps d'extraction avec la croissance des seuils de connexité et la décroissance du seuil de support. Cette réduction peut atteindre des valeurs jusqu'à 16% pour $s = 2$, 22% pour $s = 3$ et 27% pour $s = 4$. Ces réductions sont la conséquence de la réduction du nombre de motifs visités pour tester la condition de la contrainte anti-monotone (Figure 6.23). La réduction du nombre de motifs visités a les mêmes tendances de variation que la réduction du temps d'extraction avec les seuils de connexité, ($\kappa = \mu$), et de support relatif, σ_{rel}, une croissance avec la connexité et une décroissance avec le support.

Le régime d'opération avec la contrainte sur CRSM implémentée activement et avec le filtrage de la contrainte sur CM, ($\kappa = \mu$) assure une optimisation de point de vue du temps d'extraction et des nombres de motifs visités et extraits.

6.2.6 Extraction avec la conjonction de contraintes sur CRSM et CM ($\mu > \kappa$)

La conjonction des contraintes sur CRSM et sur CM peut tirer profit des fortes réductions du nombre de motifs testés et du temps d'extraction dans le cas de grandes valeurs de la connexité relative au support minimum. Dans le cas impliquant la relaxation de la contrainte sur CM, décrit dans la sous-section précédente, les valeurs des seuils des contraintes sur CRSM et CM sont maintenues égales et inférieures à la valeur 8. De cette manière, la capacité de réduction de la contrainte sur CRSM n'est pas entièrement utilisée. En permettant à la conjonction de contraintes d'opérer aussi à des grandes valeurs de la CRSM nous pouvons obtenir des motifs connexes de fréquence maximale très efficacement. Par exemple, si la conjonction avec relaxation pouvait réduire le nombre de motifs visités seulement avec maximum 40% (pour des valeurs

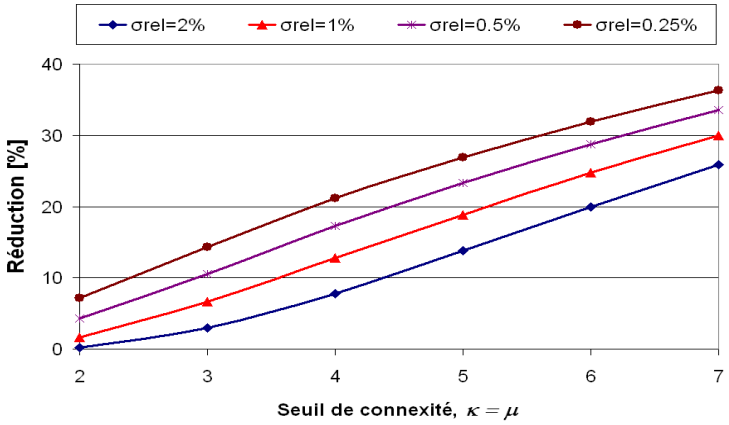


FIG. 6.23 – La réduction du nombre de motifs visités dans l'extraction CRSM+CM ($\kappa = \mu$) par rapport à celle pour CM suivant la variation des seuils de connexité et de support relatif, σ_{rel}, pour $s = 3$.

grandes des seuils de connexité mais inférieures à la valeur 8), maintenant on peut voir dans la Figure 6.24 qu'on peut atteindre des valeurs plus élevées de cette réduction.

Le graphique illustre les réductions du nombre de motifs visités, N_{vis}, obtenues pour une conjonction des contraintes de connexité en utilisant un seuil μ variable et différent de κ et un seuil κ fixé. La réduction du N_{vis} croît avec le seuil de la contrainte sur CRSM et est presque indépendant du seuil de support relatif (la conséquence de la présence d'objets grands et connexes dans la scène étudiée).

On observe une inversion à proximité de $\mu = 16$ signifiant que pour des valeurs grandes du seuil de CRSM, un seuil de support élevé assure une réduction plus accentuée. La réduction a un comportement similaire vis-à-vis du nombre de symboles utilisé, s, c'est-à-dire que, pour des valeurs grandes du seuil de CRSM, la réduction est plus marquée pour des valeurs grandes de s.

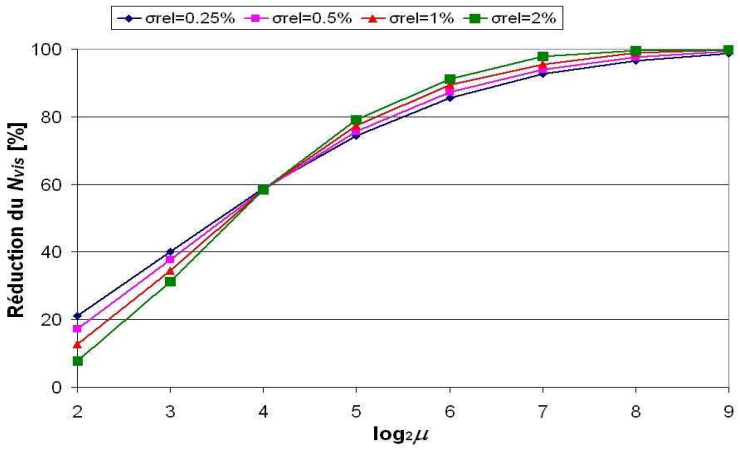
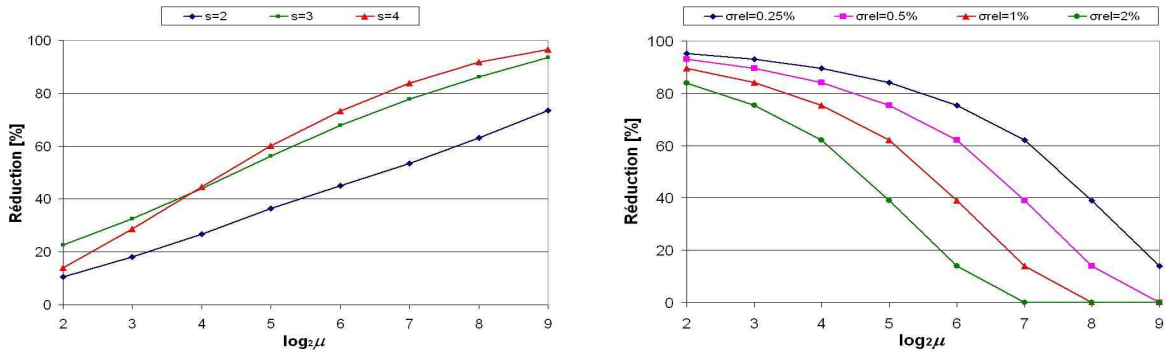


FIG. 6.24 – La réduction du nombre de motifs visités dans l'extraction avec la conjonction de contraintes sur CRSM+CM ($\mu > \kappa$) par rapport à celle avec seulement CM en fonction du seuil μ de la CRSM et du seuil de support σ_{rel} pour un seuil de CM fixe, $\kappa = 6$ et $s = 3$

Par conséquent, les temps d'extractions bénéficient de la réduction présentée dans la Figure 6.25a). Le comportement de la réduction du temps d'extraction est celui attendu : une

augmentation avec les croissances de μ et de s.



a) b)

FIG. 6.25 – a) La réduction du temps d'extraction de MSFG avec CRSM+CM ($\mu > \kappa$) en fonction du seuil de contrainte sur CRSM et le nombre de symboles pour un seuil de contrainte sur CM fixe, $\kappa = 6$ et $\sigma_{rel} = 0.5\%$ et b) La réduction du nombre de MSFG extraits avec la conjonction de contraintes sur CRSM+CM ($\mu > \kappa$) par rapport à la contrainte sur CRSM en fonction de seuils σ_{rel} et μ, dans le cas $s = 2$ et $\kappa = 6$.

La Figure 6.25b) montre la réduction du nombre de motifs due au filtrage avec la contrainte sur CM de motifs extraits avec la contrainte sur CRSM, une réduction qui décroît avec l'augmentation des seuils de CRSM et de CS.

Un problème de ce type d'extraction peut être constitué par la réduction excessive du nombre de motifs extraits dans le cas d'une croissance exagérée du seuil de CRSM. Comme le nombre de motifs extraits peut devenir moindre que le nombre de motifs obtenu avec l'utilisation seulement de contrainte sur CM, l'extraction peut être considérée incomplète par rapport à la contrainte sur CM.

Le cas concret d'une extraction avec la conjonction des contraintes de connexité discutées pour les paramètres $s = 2$ et $\kappa = 6$ est présenté ici, extraction qui révèle une caractéristique de cette Base de Données ADAM. Les valeurs des nombres de motifs extraits et des temps d'extraction suivant la variation du seuil de support relatif, σ_{rel} et du seuil de la contrainte sur CRSM, μ, sont données dans le Tableau 6.3.

	$\sigma_{rel} = 0.25\%$		$\sigma_{rel} = 0.5\%$		$\sigma_{rel} = 1\%$		$\sigma_{rel} = 2\%$	
μ	$t_{ex}[s]$	N_{motifs}	$t_{ex}[s]$	N_{motifs}	$t_{ex}[s]$	N_{motifs}	$t_{ex}[s]$	N_{motifs}
6	187	295	169	295	150	295	142	295
8	181	295	162	295	142	295	126	295
16	163	295	145	295	126	295	108	292
32	145	295	127	295	108	292	92	285
64	127	295	109	292	92	285	73	221
128	109	292	92	285	73	221	54	114
256	92	285	73	221	53	114	27	21
512	73	221	53	114	27	21	16	0
CM, $\kappa = 6$	219	295	198	295	177	295	152	295

TAB. 6.3 – Nombre de motifs et temps d'extraction CRSM+CM ($s = 2, \kappa = 6$)

On peut voir, dans la dernière ligne, que le nombre de motifs extraits avec seulement la contrainte sur CM pour $s = 2$, $\kappa = 6$ et pour toutes les valeurs étudiées du seuil de support σ_{rel} est de 295, fait qui donne une idée sur la grandeur et le degré de connexité des zones couvertes

par ces motifs. Les valeurs des temps d'extraction et des nombres de motifs extraits avec la conjonction avec relaxation sont données dans la ligne correspondante à $\mu = 6$. On observe que l'augmentation contrôlée du seuil μ peut diminuer considérablement le temps d'extraction en préservant le nombre de motifs. Dans le tableau, les plus basses valeurs du nombre de motifs de 295 (obtenu pour différents seuils σ_{rel}) s'alignent en diagonale du tableau et ces 295 de motifs ont une connexité globale, $CG \geq 160000$.

Avec les données du tableau est réalisée la Figure 6.26 qui illustre l'importance de la grandeur de connexité globale, CG, mentionnée antérieurement comme une mesure anti-monotone.

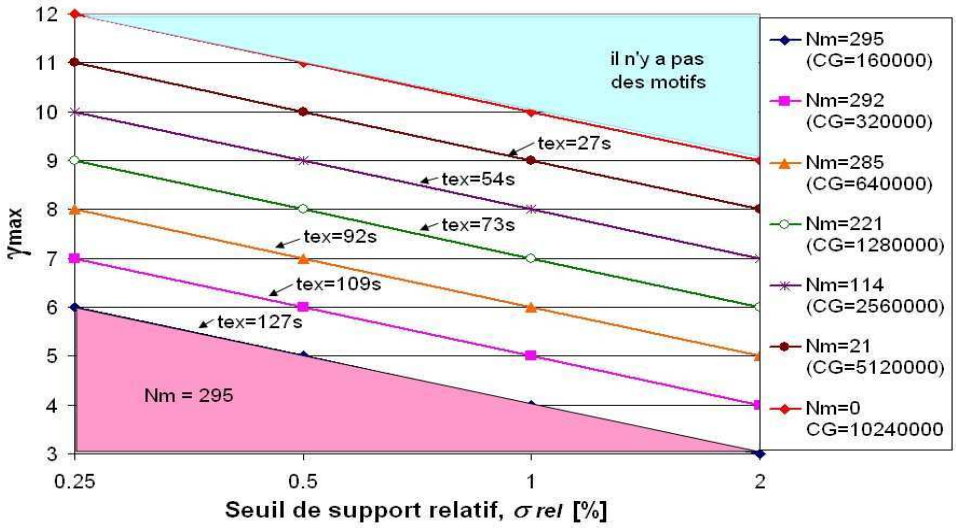


FIG. 6.26 – Les frontières dans l'espace $\gamma_{max} \times \sigma_{rel}$ des zones avec les mêmes nombre de MSFG extraits et connexité globale, CG. ($\gamma_{max} = log_2\mu$)

Les lignes inclinées du graphique constituent de frontières supérieures pour les zones de l'espace qui supportent des valeurs constantes du nombre de motifs extraits, de la connexité globale et de la valeur du temps d'extraction. Par exemple, la zone violette au-dessous de la ligne correspondant à $N_m = 295$ et/ou $CG = 160000$ assure un nombre de 295 motifs. Si l'extraction est réalisée en un point de cette frontière le temps d'extraction se réduit à la valeur minimale de 127 secondes au lieu des valeurs du temps de l'extraction avec la conjonction de contraintes sur CRSM+CM avec relaxation (187 secondes pour le même nombre de motifs). De point de vue du temps, une extraction effectuée avec des paramètres correspondants aux points situés sous la frontière conduit à un temps d'extraction plus long que sur la frontière.

En utilisant l'extraction avec la conjonction CRSM+CM ($\mu > \kappa$) peuvent être obtenus directement les motifs connexes (avec un seuil de CM, κ) les plus fréquentes par l'établissement d'une valeur grande pour le seuil de la connexité relative au support minimum, μ.

Selon les définitions 4.5, 4.7 et 4.10 on a $CG = \sigma \times CRSM = support \times CM$. D'ici on obtient $CRSM = CM \times support/\sigma = CM \times \gamma$, où γ est la sur-couverture.

Pour un seuil de CM, $\kappa = 6$, la possibilité de variation de la CM est bornée dans l'intervalle $[6, 8)$. Ainsi, une grande valeur du seuil μ de CRSM peut être obtenue seulement si le support est très grand en comparaison avec le seuil σ, autrement dit si la sur-couverture $support/\sigma$ des motifs est grande. De cette manière, sont obtenus les motifs qui maximisent le critère de fréquence plus efficacement qu'en cherchant parmi les résultats de l'extraction avec la contrainte

sur CM.

De grandes valeurs de la CRSM d'un motif impliquent aussi des grandes valeurs de sa fréquence. Et l'existence des motifs très connexes qui couvrent de grandes zones signifie en général l'intervention humaine ou des propriétés spécifiques de la croûte terrestre dans la scène observée.

6.3 Résultats qualitatifs et interprétations

Dans la section précédente est étudié le réglage des paramètres s, σ_{rel} et κ (or μ) de point de vue du nombre total de motifs, N_m, de leurs distributions en fonction de la longueur et du temps d'extraction. La démarche s'est intéressée aux motifs séquentiels, MS, aux motifs séquentiels fréquents, MSF, et aux motifs séquentiels fréquents groupés, MSFG.

Le nombre de MS croît avec le nombre de symboles, s, pour toutes les longueurs comme il était attendu. Les maximums de la distribution $N_m(L)$ se déplacent vers des L petits avec l'augmentation de s. Une conséquence des ces dépendances est une plage réduite de variations pour $N_m(20)$ en fonction de s petits : des valeurs comprises entre 3×10^4 pour $s = 2$ et 4×10^5 pour $s = 4$.

Le nombre de MSF a une dépendance normale avec la variation de s et du seuil σ_{rel} - c'est-à-dire il croît avec l'augmentation de s et la diminution de σ_{rel}. Une dépendance similaire a le temps d'extraction. La distribution de $N_m(L)$ présente de maximums qui se déplacent vers des longueurs petites avec la croissance de s et σ_{rel}. La conséquence, pour des valeurs petites de s, est une inversion $N_m(20)_{s=2} > N_m(20)_{s=3}$, un résultat intéressant pour l'utilisateur. Au passage de MS à MSF la réduction de N_m est forte, de cent fois, mais le nombre de motifs longs se réduit d'avantage, plus de mille fois (voir le Tableau 6.7).

Dans le cas de l'extraction de MSFG, en comparaison avec le cas de MSF, N_m se réduit avec la croissance des mesures de connexité, soient-elles κ, μ ou κ_G (par exemple d'environ dix fois pour $\kappa = 5$). Le nombre de motifs longs diminue aussi, mais pas dans la même mesure. Pour des valeurs très grandes des seuils κ et μ se produit une inversion de plus, $N_m(s = 2, k > 6) > N_m(s = 3, k > 6)$, un cas dans lequel la quantification réduite aide à la connexité.

Concernant les valeurs des variables principales (σ, s et κ), on considère que le seuil du support σ, peut être choisi par l'utilisateur en fonction de la dimension la plus petite de zone de la scène désirée à être détectée. Dans cette STIS il y a des objets très étendus et les dépendances des paramètres étudiés en fonction du support sont généralement très faibles. Ainsi la valeur de σ peut être dictée par d'autres considérations, comme le temps d'extraction par exemple. Les nécessités impliquées par nos objectifs spécifiques (N_m petit, L grand, t_{ex} petit) conduisent jusqu'à maintenant au choix de petites valeurs de s et à des grandes valeurs de κ. Il reste à voir quelles sont les exigences imposées par l'objectif lié à la grande couverture des motifs en pixels de la scène, le cinquième objectif (N_C grand).

6.3.1 Stratégies de sélection des motifs

Dans la dernière section, il est montré que la méthode proposée améliore l'efficacité du processus d'exploration, principalement en réduisant le nombre de motifs qui sont découverts. Toutefois, nous n'avons pas discuté de la qualité des motifs découverts. La principale raison de l'absence de cette discussion jusqu'à ce point est l'inexistence d'une évaluation exacte de la qualité des motifs découverts. La seule façon de faire une telle évaluation est de comparer les

motifs découverts avec les connaissances existantes du domaine d'application, qui doivent être assez bonnes pour distinguer les motifs pertinents de non pertinents.

6.3.1.1 La couverture des pixels de la scène avec les motifs extraits

Un premier type d'évaluation des motifs extraits peut se faire de point de vue de leur couverture en pixels de la scène du projet ADAM. Le Tableau 6.4 donne quelques informations dans ce sens en faisant aussi des comparaisons entre les types d'extraction développés. N_m est le nombre total de motifs extraits, N_L est le nombre de motifs de longueur L, N_{CP18} est le nombre de pixels couverts par un seul 18-motif (pixels purs) exprimé en pourcents relatif à la scène entière et N_{CT18} est le pourcentage des pixels couverts par tous les 18-motifs relatif à la scène.

Type de motif	s	σ_{rel} [%]	κ	N_m	N_{18}	N_{19}	N_{20}	N_{CP18} [%]	N_{CT18} [%]
MS	2			510 027	133 464	97 033	32 024		
MSF	2	0,5		7 926	144	64	27		
MSFG (B)	2	0,5	5	681	24	15	7	18,79	72,63
MSFG	2	0,5	5,5	484	12	10	2	21,84	65,22
MSFG (C)	2	0,5	6	295	10	1	1	23,44	63,58
MSF	2	1,0		4 647	65	35	19		
MSFG	2	1,0	5	666	21	15	7	18,33	71,81
MS	3			10 367 679	1 344 331	673 781	181 151		
MSF	3	0,5		43 814	157	60	14		
MSFG	3	0,5	5	2 338	38	14	2	11,04	32,36
MSFG (A)	3	0,5	5,5	1 104	14	6	2	11,09	19,54
MSFG	3	0,5	6	479	6	2	1	3,36	5,33
MSF	3	1,0		23 038	68	27	3		
MSFG	3	1,0	5	2 160	28	8	1	11,38	30,77

TAB. 6.4 – La comparaison des nombres de MS, MSF et MSFG et des couvertures de la scène avec des 18-motifs (IVDN).

On définit trois points de fonctionnement qui visent des différents objectifs d'intérêt [119] :
- Le point de fonctionnement (A) qui maximise la contribution des pixels purs dans le nombre total de pixel couverts par de 18-motifs ; $(N_{CP18}/N_{CT18})_{max} = 56,55\%$;
- Le point de fonctionnement (B) qui assure la plus grande couverture totale avec 18-motifs, $N_{CT18} = 72,63\%$ et aussi le nombre maximal de motifs complets, $N_{20max} = 7$;
- Le point de fonctionnement (C) qui maximise la couverture avec 18-motifs purs $N_{CP18max} = 23,44\%$.

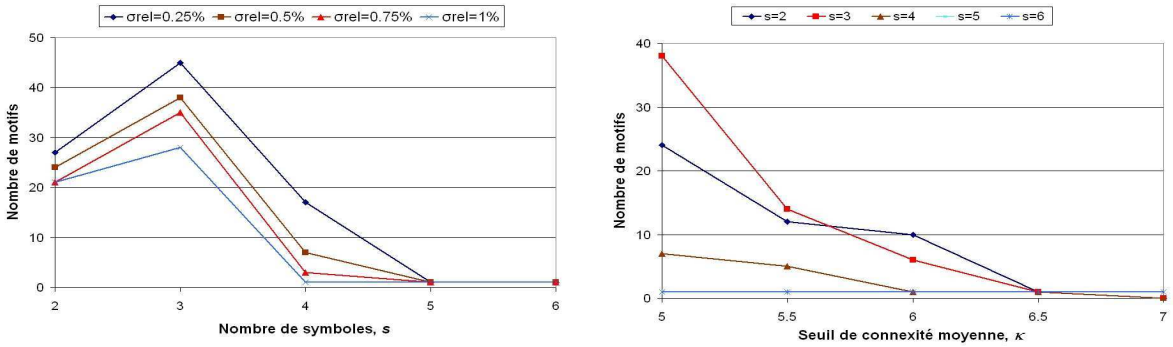
Le Tableau 6.4 montre la réduction successive du nombre de motifs en appliquant les seuils de support et de connexité moyenne.

Pour $s = 2$, la croissance de κ produit l'augmentation de N_{CP18} bien que N_{CT18} diminue. Les pixels des cultures très connexes décrits par le plus petit nombre de symboles ont une chance en plus d'être purs. Quand s croît, par exemple pour $s = 3$, la croissance de κ ne produit pas le même effet, la densité en motifs décroît et le seuil de fréquence elimine des motifs.

On arrive à considérer les motifs de longueur 18 et 19 parce que le nombre de motifs de longueur maximale, $L = 20$, est petit et la couverture de la vérité terrain assurée par les 20-motifs est faible. Certaines acquisitions souffrent des perturbations atmosphériques et des conditions du capteur. De plus, selon les zones qui sont considérées, les cycles phénologiques d'un type donné de culture ne démarrent pas et ne finissent pas toujours à la même date, étant donné que

des différentes conditions pédologiques, de fertilisation et d'irrigation sont présentes. Il est donc impossible de faire appel aux MSFG ayant autant d'événements que le nombre d'acquisitions, c'est-à-dire 20-MSFG, pour décrire correctement les cycles phénologiques. Ainsi, on se concentre sur des motifs longs incomplets, par exemple les 18 et 19-MSFG tels qu'ils soient suffisamment généraux pour envisager l'apparition éventuelle à des différentes dates et qui ignorent une ou deux valeurs bruitées.

Les dépendances du nombre de 18-MSFG des paramètres s, σ_{rel} et κ présentées dans la Figure 6.27 a) et b) montrent deux comportements intéressants : un maximum pour $s = 3$ et $\kappa = 5$ et un inversion $s = 2/s = 3$ pour $\kappa = 6$. Pour des degrés élevés de connexité moyenne les cas $s = 2$ et $s = 3$ sont presque similaires.



a) b)

FIG. 6.27 – a) Le nombre de 18-MSFG en fonction du nombre de symboles s et le seuil de support relatif, σ_{rel}, pour $\kappa = 5$ et b) Le nombre de 18-MSFG en fonction du seuil de connexité moyenne, κ, du nombre de symboles s, pour $\sigma_{rel} = 0,5\%$.

Un de nos objectifs spécifiques a été d'avoir une couverture convenable des pixels de la scène avec les motifs extraits. Pour les motifs de longueur 18, les pourcentages des pixels couverts relatif à la scène entière sont présentés dans la Figure 6.28.

Les pourcentages de la couverture avec les 18-MSFG diminuent extrêmement avec l'augmentation du nombre de symboles, s, de sorte que seulement les valeurs $s = 2$ et $s = 3$ peuvent être considérées d'intérêt. Comme cela est attendu, ces pourcentages descendent avec la croissance du degré de connexité par suite de la décroissance du nombre de motifs. La variation du seuil du support relatif, σ_{rel}, dans la gamme $0,25\% - 2\%$, a une influence insignifiante sur le degré de couverture. Le maximum de la couverture, $72,63\%$, est obtenu pour $s = 2$ et $k = 5$, valeurs qui définissent le point de fonctionnement B (montré dans la Figure 6.28), point qui a été énoncé dans le Tableau 6.4.

Comme la série temporelle a 20 images, un motif de longueur 18 laisse deux dates d'acquisition non considérées. Dans cette situation, un pixel peut être couvert par différents 18-motifs. Un pixel qui est couvert par un seul 18-motif est dénommé un *pixel pur*. Evidemment, les pixels purs acquièrent une importance spéciale dans notre démarche de caractérisation des cultures agricoles par les évolutions des pixels. La Figure 6.29 a) et b) présentent les comportements des pourcentages de couverture par des pixels purs en fonction de la variation des paramètres s, σ_{rel} et κ. Pratiquement, la variation de σ_{rel} a une influence très faible sur la couverture de pixels (Figure 6.29a). Le point de fonctionnement A, figuré dans le graphique, correspondant à un maximum ici, a une signification particulière parce que pour les paramètres $s = 3$ et $\kappa = 5,5$ le rapport N_{CP18}/N_{C18} atteint $56,55\%$ (le poids maximum de pixels purs).

Un autre comportement intéressant est présenté dans la Figure 6.29b), une augmentation nette du nombre de pixels purs avec la croissance du degré de connexité, pour $s = 2$ dans

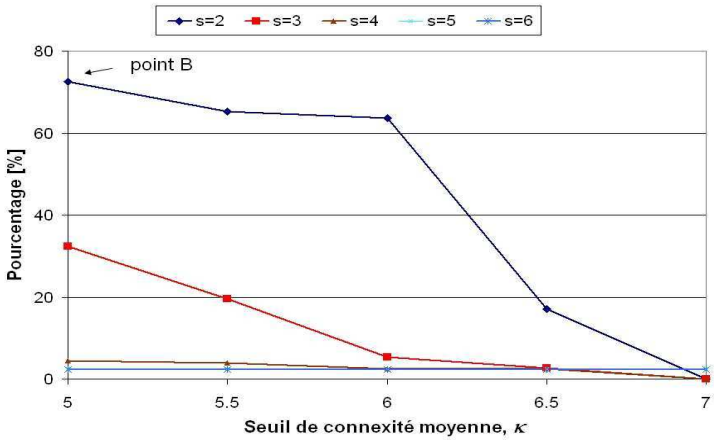
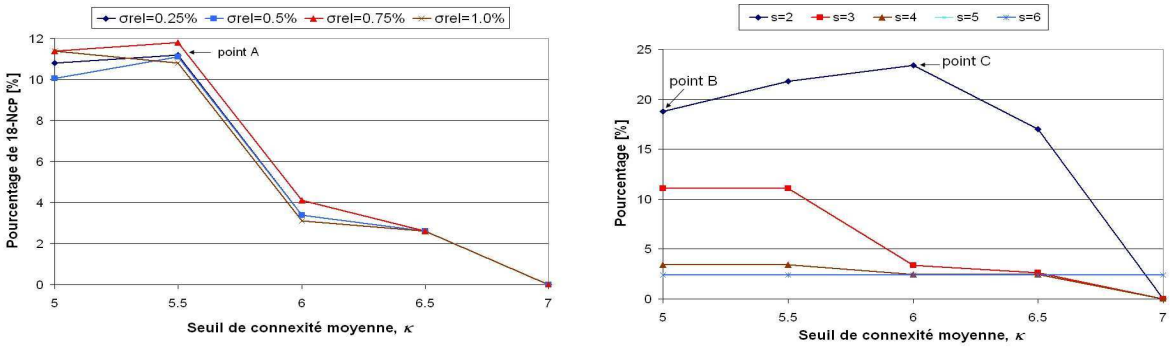


FIG. 6.28 – Le pourcentage de couverture avec les 18-MSFG en fonction du seuil de connexité moyenne, κ, et du nombre de symboles, s, pour $\sigma_{rel} = 0,5\%$.



a) b)

FIG. 6.29 – a) Le pourcentage de pixels purs couverts par les 18-MSFG en fonction du seuil de connexité moyenne, κ, et du seuil du support relatif σ_{rel}, pour $s = 3$ et b) Le pourcentage de pixels purs couverts par les 18-MSFG en fonction du seuil de connexité moyenne, κ, et du nombre de symboles, s, pour $\sigma_{rel} = 0,5\%$.

la gamme $\kappa \in [5, 6]$, en dépit de la diminution du nombre de motifs. C'est un argument fort pour considérer le point C ($\kappa = 6, s = 2$) un point de fonctionnement d'intérêt. En effet, la binarisation des valeurs des pixels peut donner de très bons résultats concernant l'efficacité et l'interprétation de l'extraction des motifs.

6.3.1.2 L'utilisation d'une Vérité Terrain de la scène

Dans l'étude de la couverture végétale de la scène du projet ADAM et de la vérité terrain utilisée sont introduits des paramètres caractéristiques pour mesurer la contribution des motifs dans la caractérisation complète et correcte des cultures agricoles. Ainsi, les suivantes définitions contextuelles sont nécessaires.

- ***Pixel pur*** - pixel couvert par un seul motif.
- ***Motif pur*** - motif avec une pureté globale maximale (près de 100%). Un motif complet pur appartient à une seule culture ; il est mono-culture. Un motif incomplet qui a une pureté très grande peut être mono-culture ou pluriculture. Dans ce dernier cas, les variantes temporelles du motif permettent de distinguer les cultures.
- ***Motif complet*** - motif de longueur maximale (L = nombre d'images) pas obligatoirement

pur. Les motifs ayant la longueur inférieure à la longueur maximale sont définis comme *incomplets*.

- ***la Couverture de la Vérité Terrain, CVT***, qui est le rapport entre le nombre de pixels qui se retrouvent dans la vérité terrain et sont couverts par le motif et le nombre de pixels de celle-ci.
- ***la Culture Principale, CP***, est la culture qui correspond à la majorité de pixels couverts par un motif ou par une variante temporelle d'un motif.
- ***la Couverture de la Culture Principale*** dans la vérité terrain, ***CCP***, qui caractérise la contribution du motif dans la description correcte de sa culture principale dans la vérité terrain. La culture principale ou majoritaire d'un motif est la culture qui bénéficie du nombre le plus grand des pixels couverts par le motif. La CCP est donnée par le rapport entre le nombre de pixels couverts par le motif qui appartiennent à la culture définie comme principale pour ce motif et le nombre de pixels pour cette culture dans la vérité terrain.
- ***la Pureté, P***, peut être définie pour un motif complet ou pour une variante temporelle d'un motif comme le pourcentage des pixels couverts par la culture principale par rapport aux pixels couverts par le motif.
- ***la Pureté Globale, PG***, peut être définie pour un motif incomplet qui a plusieurs variantes temporelles, comme le pourcentage de la somme des pixels couverts par les cultures définies comme principales dans chaque variante par rapport au nombre total de pixels couverts par le motif.

La vérité terrain pour la zone Fundulea (Progresu 1 - 2 et Tipei ; l'année 2001) a été obtenue de l'Institut de Recherche et Développement en Agriculture et est présentée dans la Figure 6.30. La carte représente $5,9\%$ de la surface de la zone étudiée mais elle contient toutes les cultures importantes de la scène entière. Les zones blanches intérieures de la carte correspondent à des forêts diverses.

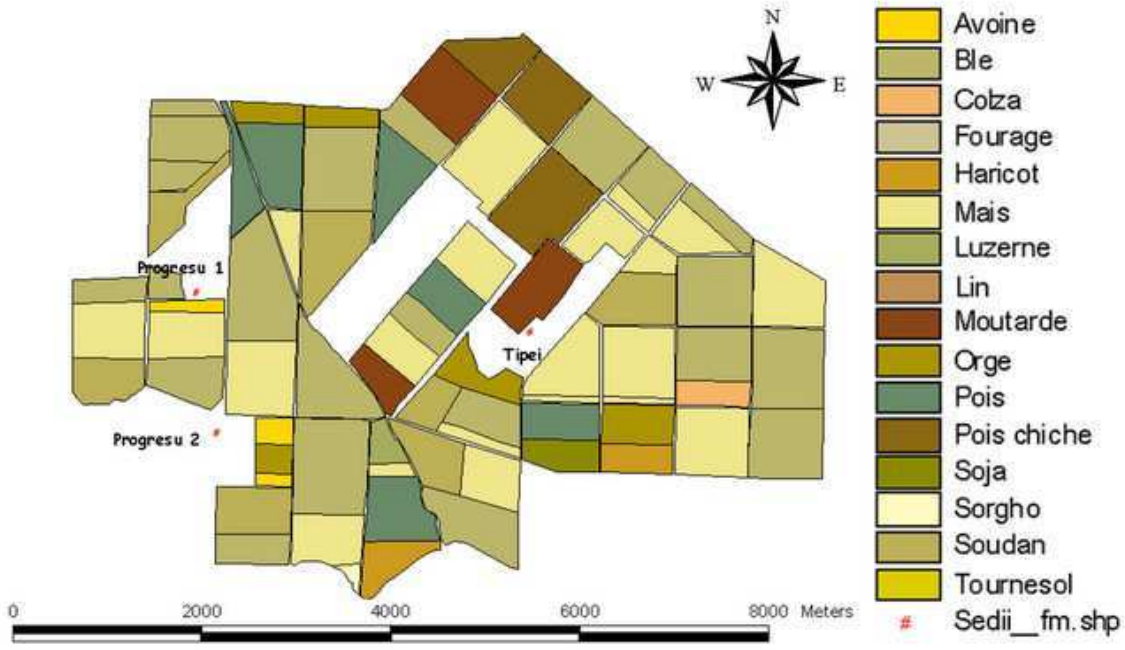


FIG. 6.30 – La vérité terrain de la zone Progresul 1 - 2 et Tipei pour l'année 2001

Avec une vérité terrain, il est possible de faire correspondre, temporellement et spatialement, les MSFG avec les types connus de cultures. Plus précisément, les pixels couverts par un motif donné α sont divisés en sous-ensembles, (variantes du motif), chaque sous-ensemble étant lié à

une distribution donnée de dates des événements. Le nombre de pixels d'un tel sous-ensemble, pour une distribution donnée de dates d'occurrence, est indiqué par $cov(\alpha, i)$. Une culture principale est ensuite affectée à chaque sous-ensemble conformément à la vérité terrain (chaque pixel couvert vote pour la culture à laquelle il correspond). Au sein d'un tel sous-ensemble, tous les pixels correspondant à cette culture dominante sont appelés pixels dominants. Ils sont signalés par $d(cov(\alpha, i))$. Ensuite, une pureté globale, $PG(\alpha)$, est calculée. Elle est inspirée par des mesures de pureté qui sont utilisées pour évaluer la pureté globale d'un cluster [207]. Plus formellement, si D est l'ensemble de toutes les distributions observées des dates d'occurrence, $PG(\alpha)$ définie au-dessus devient :

$$PG(\alpha) = \frac{\sum_{i \in D} d(cov(\alpha, i))}{cov(\alpha)} \tag{6.3}$$

La manière abrégée de caractériser un motif est : s_1xm_1_s_2xm_2_s_3xm_3 etc. (s =; SR =; CM =) où l'évolution est décrite par s_i qui sont les symboles et m_i qui sont leurs multiplicités écrites dans leur ordre d'occurrence ; entre parenthèses sont donnés les paramètres s, le nombre de symboles, SR, le support relatif, CM, la connexité moyenne auxquels on peut ajouter les paramètres définis dans cette section CVT, CCP et PG.

Le tableau 6.5 présente la modalité de calcul de la pureté globale du MSFG 1x14_2x4 (IVDN, $SR = 17,86\%$, $CM = 6,3$) extrait avec $s = 2$ symboles, motif qui assure une bonne couverture de la vérité terrain, $CVT = 30,42\%$ et la meilleure couverture de la culture principale $CCP = 75,59\%$. Usuellement, des telles couvertures élevées correspondent à une pureté globale pas très grande (ici, $PG = 76,36\%$). La culture principale est le maïs et les cultures secondaires le petit pois, l'herbe du Soudan et le soja. Une première action est d'établir les variantes temporelles du motif. Le motif ayant la longueur 18 pour une série de 20 images il y a $C_{20}^{18} = 190$ possibilités des variantes temporelles. Dans ce cas, 13 variantes temporelles sont trouvées et codées en binaire. Les 20 dates sont alignées avec la première date à droite. Si la variante est présente à une date donnée on code «1», si elle est absente on code «0». De cette manière, on obtient le code de la variante temporelle. Par exemple, la première variante a le code binaire 00111111111111111111 (262.143 en décimal). Ça signifie que le 18-motif couvre les premières 18 dates et les dates manquantes sont la 19-ème et la 20-ème, fait consigné dans le tableau. À l'aide de ce code on fait la localisation de la discrimination temporelle d'un motif (voir l'annexe B.2). Dans le tableau les codes binaires sont donnés seulement pour les variantes bien peuplés. Après l'établissement des variantes temporelles, on cherche le nombre de pixels qui couvrent les cultures de la vérité terrain pour chaque variante. Maintenant la culture principale de la variante et la pureté de la description de cette culture peuvent être établies. Ainsi, la première variante qui couvre 4223 pixels dans la VT, décrit correctement avec 2156 pixels la culture de l'herbe du Soudan définie comme principale et la pureté obtenue est $48,75\%$. Cette valeur réduite est donnée par les populations comparables des pixels qui couvrent les autres cultures : maïs et petit pois. Les 12 variantes qui restent ont le maïs comme culture principale et 7 d'entre elles ont plus de 10 pixels et sont présentées dans le tableau. En conséquence, les MSFG peuvent être évalués qualitativement pour offrir à l'utilisateur des informations sur la description correcte de la scène et sur la possibilité de sélectionner de bons candidats pour un éventuel clustering.

Variante du motif		1	2	5	6	7	8	9	10	13	Pixels (%)	CVT (%)	
Pixels		4 423	1 568	2 716	4	59	5 127	3 721	35	303	17,96	30,42	
Code binaire		262 143	507 903	524 286			999 423	1 015 806		1 048 565			
Dates manquantes		19, 20	15, 20	1, 20			15, 16	1, 16		2, 4			
Vérité Terrain											Motif		
Culture	Pixels										Pixels	CC (%)	PCM (%)
pois chiche	3 583			372				62	16	8	458	12,78	2,55
moutarde	2 849			1			10	7		3	21	0,74	0,12
blé	20 717	94	48	41			75	27		10	295	1,42	1,64
maïs	15 615	285	**1 278**	**2 141**	**3**	**59**	**4 444**	**3 321**	**18**	**254**	**11 804**	**75,59**	**65,72**
petit pois	5 023	1 825	94	15			55	48		3	2 040	40,61	11,36
orge	2 261	22	6				10	19			57	2,52	0,32
soudan	5 881	**2 156**	66	47	1		41	160		21	2 493	42,39	13,88
avoine	541	4	11				15				30	5,55	0,17
colza	470			4				25	1	2	32	6,81	0,18
lucerne	288			27			2	28			57	19,79	0,32
soja	567	1	64	6			459	2		1	535	94,36	2,98
haricot	1 167		1	62			16	22		1	102	8,74	0,57
millet	80										36	45,00	0,20
Pureté (%)		48,75	81,51	78,83	75,00	100,00	86,68	89,25	51,43	83,83			
Pureté globale (%)		**76,36**											

TAB. 6.5 – Le calcul de la pureté pour le 18-motif 1x14_2x4 ($CM = 6,3$; $SR = 17,86\%$) ayant le maïs comme culture principale

6.3.1.3 Le choix du canal spectral

Les images SPOT sont obtenues dans 3 bandes spectrales : vert, rouge et proche infrarouge.

La région verte (500 - 590 nm) donne peu de détails sur la végétation parce qu'elle correspond à l'absorption de la chlorophylle par la végétation en bonne santé. Cette bande est utile pour les détails cartographiques tels que la profondeur ou les sédiments dans les plans d'eau. Les caractéristiques telles que les routes et les bâtiments apparaissent également bien dans cette bande. On peut voir les villages mis en évidence dans les Figures C.7a) et b) de l'annexe C.

Dans la région rouge du spectre (610 - 690 nm), la chlorophylle absorbe ces longueurs d'onde dans la végétation saine. Par conséquent, cette bande est utile pour distinguer certaines espèces de plantes, ainsi que les frontières géologiques et du sol. La Figure C.8 de l'annexe C montre également des routes et des villages.

Le proche infrarouge, (780 - 890 nm), correspond à la région du spectre électromagnétique qui est particulièrement sensible à la biomasse de la végétation qui varie. Il insiste également sur la frontière sol - cultures et terre - eau et il est utilisé pour la discrimination de la végétation, pénétrant la brume.

L'IVDN compense largement le changement des conditions d'éclairage, la pente de la surface, et les différents angles de vue. Les nuages, l'eau et la neige donnent des valeurs négatives en raison d'une réflectance dans le rouge plus grande que dans le PIR. Les valeurs de IVDN pour les roches et les sols nus secs sont proches de zéro en raison de leurs réflectances semblables dans les deux bandes. En appliquant une translation des valeurs de pixels on obtient seulement des valeurs positives, comme dans ce travail. De cette manière l'eau correspond aux valeurs positives très petites, près de zéro et les pixels correspondants au sol ont des valeurs plus élevées.

Pour les données de la STIS ADAM, restent en compétition le PIR et l'IVDN. Une comparaison entre les nombres de motifs et des pixels couverts pour les deux canaux, PIR et IVDN, est présentée dans le Tableau 6.6.

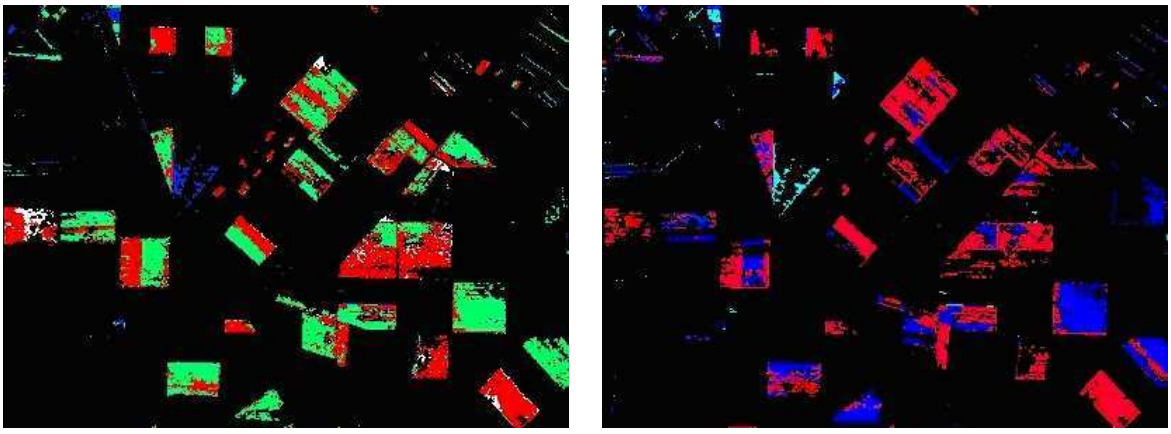
No.	Point	Bande	$\sigma_{rel}[\%]$	s	κ	N_{18}	N_{19}	N_{20}	$N_{CP18}[\%]$	$N_{C18}[\%]$
1	A	IVDN	0,5	3	5,5	14	6	2	11,09	19,54
2	A	PIR	0,5	3	5,5	12	5	1	10,08	14,20
3	B	IVDN	0,5	2	5	24	15	7	18,79	72,63
4	B	PIR	0,5	2	5	30	17	6	19,87	71,56
5	C	IVDN	0,5	2	6	10	1	1	23,44	63,59
6	C	PIR	0,5	2	6	4	1	1	22,79	35,66

TAB. 6.6 – Comparaisons IVDN - PIR pour les points d'opération A, B et C.

La signification des points d'opération reste la même : le point A assure la meilleure contribution des pixels purs, B la plus grande couverture et C le plus grand nombre de pixels purs. Pour la scène entière la couverture assurée par les 18-motifs est plus grande pour la bande IVDN dans tous les trois points d'opération, sans tenir compte de la relation entre les nombres de motifs extraits avec ces deux bandes. Même la correspondance motif - culture agricole reste la même en utilisant les deux bandes. La qualité des motifs obtenus avec la bande IVDN est, dans tous les cas, supérieure. Cela justifie le choix de ce canal de données pour extraire des MSFG pour la caractérisation de la STIS ADAM.

En général, l'utilisation du canal IVDN améliore les attributs des motifs extraits. Dans la Figure 6.31 sont présentés les motifs 1x14_2_3x3 extraits des données IVDN et PIR et qui correspondent au maïs. La comparaison entre les paramètres de caractérisation montre la qualité supérieure du motif obtenu avec l'IVDN. Pour l'utilisation de la bande IVDN sont obtenues les

valeurs : $CVT = 19,58\%$, $CCP = 66,75\%$ et la pureté globale $PG = 91,13\%$. Dans le cas de l'utilisation des données PIR, les valeurs des mêmes paramètres sont : $CVT = 14,12\%$, $CCP = 47,11\%$ et $PG = 89,88\%$. Les motifs extraits avec les données IVDN assurent une couverture plus grande et les contours des zones couvertes sont plus nets. Le motif mentionné prouve la possibilité de discriminer les cultures agricoles parmi les variantes du motif. La culture d'herbe du Soudan est visualisée en bleu ciel dans le motif extrait de données en PIR et en bleu marine dans le motif extrait de données en IVDN.



a) b)

FIG. 6.31 – Le 18-motif 1x14_2_3x3 extrait des données : a) IVDN et b) PIR.

6.3.2 Motifs courts

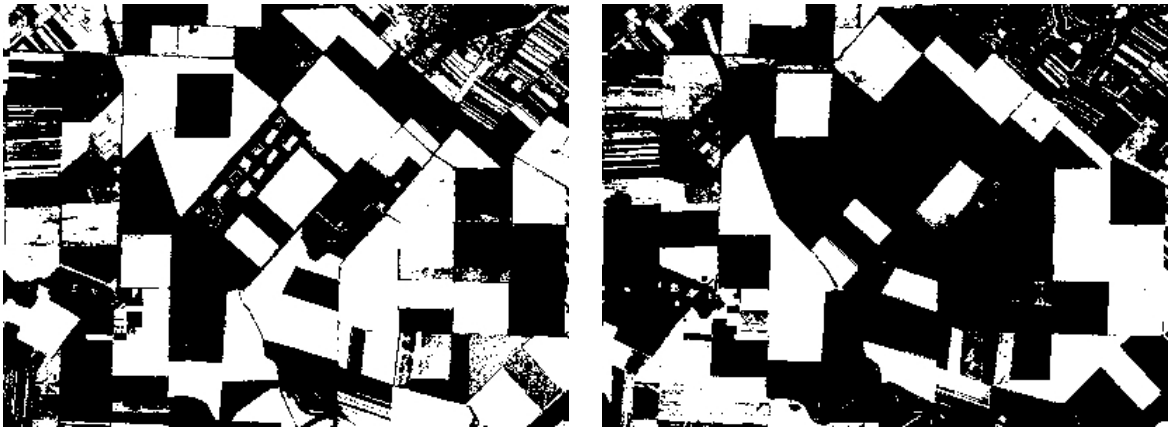
Pour expérimenter l'obtention des informations multiples sur les MSFG extraits on a considéré nécessaire d'utiliser toutes les longueurs des motifs. Les motifs courts offrent des informations avec un degré de généralité élevé. Avec la croissance de la longueur, le degré de généralité des motifs décroît mais la spécialisation, leur précision descriptive sur les évolutions des pixels, augmente. La couverture des motifs qui est grande pour les motifs courts décroît avec la croissance de leur longueur. La comparaison avec le vérité terrain offre des paramètres supplémentaires pour caractériser la qualité des motifs extraits de n'importe quelle longueur.

Un aspect important résulte de l'étude des MSFG en fonction de leur longueur : seuls les motifs courts permettent des degrés de connexité très élevés. Ainsi, on peut chercher ce type de motifs en commençant avec $\kappa = 7$. Si on obtient des attributs de discrimination des cultures agricoles pour des motifs de longueurs petites, il est possible de fusionner spatialement, avec un ordre de priorité, ces motifs qui implicitement ont une grande couverture avec des motifs longs qui offrent leur contribution de spécialisation. Ainsi, le 6-motif 3x6 ($IVDN; s = 3; SR = 54\%; CM = 7.05$; Figure 6.32) assure une discrimination entre le maïs et ses compagnons usuels dans les motifs longs et avec grande couverture, le petit pois et l'herbe du Soudan (par exemple, le 18-motif 1x14_2x4 ($s = 2; SR = 17,9\%, CM = 6,3; CVT = 30,42; PG = 76,36\%$), Figure 6.39a).

Même les motifs très courts offrent des informations utiles sur les évolutions des pixels. Par exemple, parmi les 4-MSFG extraits dans le point d'opération A, on peut trouver le 4-motif 1x2_3x2 ($IVDN; s = 3; SR = 38,05\%; CM = 6,78$). Ce motif souligne la première partie des cycles phénologiques : certaines cultures sont semées, croissent et arrivent à maturation (par exemple, cultures semées le printemps : maïs, petit pois, pois chiche et herbe du Soudan). Le motif est localisé en éclairant chaque pixel qui est couvert par le motif, tandis que les autres restent noirs. Le résultat est illustré dans la Figure 6.33a), de la zone où la vérité du ter-

FIG. 6.32 – La localisation du 6-motif 3x6 ($IVDN; s = 3$).

rain est disponible. Les régions géométriques relativement homogènes avec des frontières nettes s'affichent. Les régions blanches correspondent à différents types de champs agricoles avec des cultures tardives de printemps, tandis que les régions noires correspondent aux forêts, aux masses d'eau et à d'autres types de champs agricoles avec cultures d'automne ou précoces de printemps.



a) b)

FIG. 6.33 – La localisation du a) 4-motif 1x2_3x2 ($IVDN; s = 3$) et b) 5-motif 3x3_1x2 ($IVDN; s = 3$).

Un autre exemple d'un motif court extrait est un 5-motif séquentiels fréquents groupés (SFG), le 3x3_1x2 ($IVDN; s = 3; SR = 35,7\%; CM = 6,91$; Figure 6.33b). Il correspond aux cultures semées l'automne, particulièrement de blé.

La superposition de la localisation de ce motif et du précédent couvre plus de 95% de la zone de la scène pour laquelle la vérité terrain est disponible. Elle est présentée dans la Figure 6.34. Les pixels non affectés, en noir, représentent des forêts, des routes, des masses d'eau, des localités et un type donné de culture, à savoir celle de haricot. La zone bleue contient des pixels purs qui sont couverts par un seul des deux motifs. Elle concerne des cultures avec des cycles phénologiques de longue durée semées à l'automne ou au printemps. Les pixels couverts par les deux motifs sont colorés en rouge et correspondent aux champs d'orge et d'avoine qui ont leurs cycles phénologiques courts compris dans la période d'observation. En dépit de leurs nombre réduit d'évenements, ces motifs nous permettent en effet d'observer quatre types d'évolution. Ce genre de motif court peut donc servir pour caractériser très généralement les principales évolutions dans une STIS.

Il est possible de caractériser un motif court par les paramètres liés à la vérité terrain. Si pour les motifs très courts, comme ceux discutés au-dessus, on peut atteindre des couvertures de la vérité terrain très grandes, avec la croissance de la longueur du motif la CVT normalement décroît.

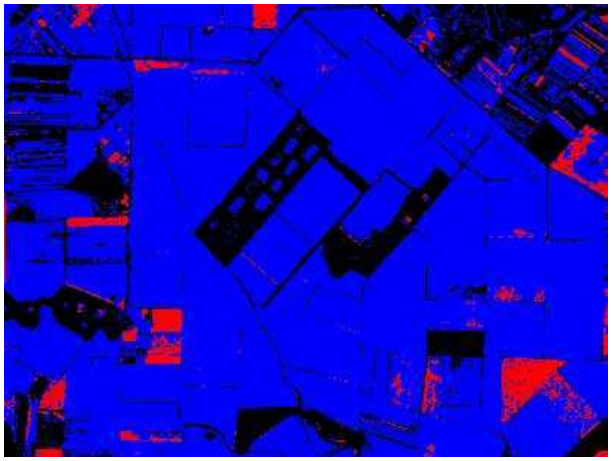
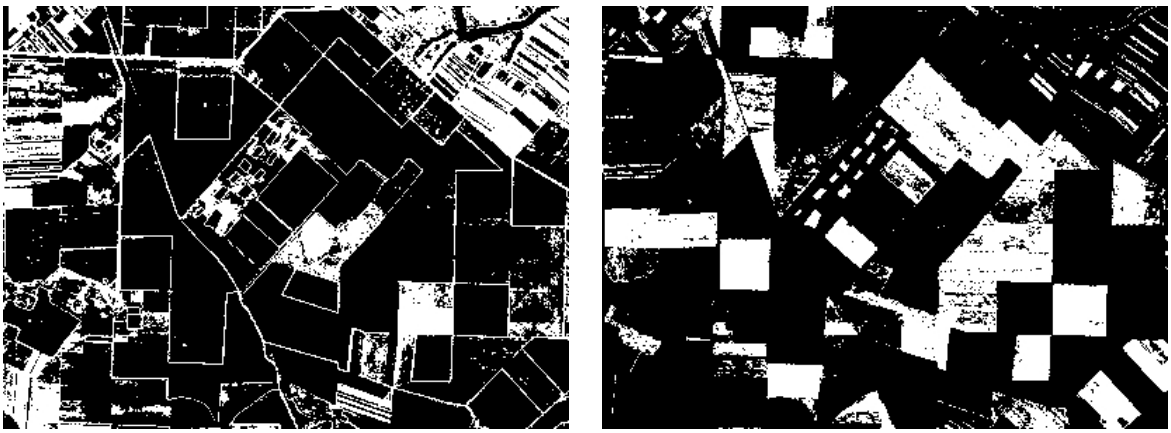


FIG. 6.34 – La superposition des motifs 1x2_3x2 et 3x3_1x2 ($IVDN; s = 3$).

6.3.3 Motifs intermédiaires

Pour la localisation des objets plus précis ou des régions, des motifs plus longs et donc plus spécifiques doivent être considérés. Au niveau suivant, les motifs intermédiaires ont des propriétés qui combinent la généralité de ceux courts avec la spécialisation de ceux très longs. Par un choix attentif, ces motifs peuvent offrir des informations très spécialisées.



a) b)

FIG. 6.35 – La localisation du a) 8-motif 2x8 ($IVDN; s = 3$) et b) 13-motif 1x12_2 ($IVDN; s = 3$).

Par exemple, le 8-motif 2x8 ($s = 3; SR = 38\%; CM = 6,2$), représenté dans la Figure 6.35a) met en évidence des villages, des zones de forêt et un champ de haricot ainsi qu'une vraie carte des routes et des bordures des champs agricoles.

Les motifs intermédiaires assurent des couvertures plus grandes que les motifs longs (voir le Tableau 6.7) mais parfois de même niveau de spécialisation que ceux-ci. Le 13-motif 1x12_2 ($IVDN; s = 3; SR = 16,3\%; CM = 6,07; CVT = 34,53\%; CCP = 88,68\%; PG = 67,92\%$)

est localisé dans la Figure 6.35b). La culture de maïs est très bien couverte, tous les champs avec cette culture sont représentés et les frontières sont très nettes. La couverture de la culture principale atteint un niveau très grand de $88,68\%$, ce fait étant une raison de plus pour l'analyse de tous les motifs extraits d'une STIS, pas seulement les maximaux ou les complets. Les autres cultures représentées sont le petit pois, le pois chiche et l'herbe du Soudan qui peuvent être partialement discriminées selon les dates d'occurrence.

a) b)

FIG. 6.36 – La localisation du a) 15-motif 2_3x10_1x4 ($IVDN; s = 3$) et b) 15-motif 2_3x11_1x3 ($IVDN; s = 3$).

De très bons candidats pour un éventuel regroupement des motifs sont les 15-motifs 2_3x10_1x4 ($SR = 9,2\%; CM = 6,1; CVT = 22,36\%; CCP = 61,38\%; PG = 96,31\%$) et 2_3x11_1x3 ($SR = 7,9\%; CM = 6,05; CVT = 21,16\%; CCP = 55,99\%; PG = 92,84\%$) qui ont comme culture principale le blé (Figure 6.36a) et b). Ces motifs ne contiennent pas des champs avec moutarde, la culture associée souvent au blé dans la majorité de motifs pour les cultures d'hiver.

6.3.4 Motifs longs

Les motifs complets, (longueur égale au nombre total d'images), offrent la plus grande spécialisation. Mais leur couverture de la vérité terrain et leur pureté sont faibles, ces motifs n'étant pas obligatoirement purs.

Même pour le plus petit nombre de symboles utilisés, $s = 2$, c'est-à-dire une binarisation des images, le nombre de MSFG complets est petit. Pour l'utilisation d'un seuil de connexité moyenne $\kappa \geq 5$ et pour un seuil de support relatif $\sigma_{rel} \geq 0,5\%$, le nombre maximal de motifs complets, 7, est obtenu pour les limites inférieures des intervalles (le point d'opération B). Malheureusement, tous les motifs complets ne sont pas aussi mono-culture (leurs pixels couverts peuvent appartenir à des diverses cultures agricoles). Ainsi, seulement les motifs 2x15_1x5, 1x16_2x4 (Figure 6.39e) et 1x20 ont une correspondance quasi univoque dans le vérité terrain : du blé ($PG = 89,58\%$), du maïs ($PG = 88,69\%$) et, respectivement l'eau et couvrent seulement $0,09\%$ de la scène. Le reste de 4 motifs correspondent à quelques cultures qui usuellement apparaissent ensemble (2x14_1x6 blé et moutarde présentés dans la Figure 6.38 ; 1x14_2x6 présentés dans la Figure 6.39d) et 1x12_2x8 petits pois et herbe du Soudan ; 2x20 forêt et haricot). Tous les 7 motifs complets extraits couvrent $22,5\%$ de la scène entière. La superposition de ces motifs est illustrée dans la Figure 6.37. Seulement quelques régions ont des frontières nettes.

Pour accroître la couverture, la connexité et parfois la pureté globale il est nécessaire d'étudier

FIG. 6.37 – La localisation de la superposition des 20-motifs SFG obtenus avec la point d'opération B.

également les motifs de longueur 18 et 19.

Ainsi, pour le point d'opération B, on obtient 24 motifs de longueur 18 et 15 motifs de longueur 19 qui peuvent offrir de bons «candidats» pour un éventuel clustering. Le Tableau 6.7 donne des informations sur les principaux MSFG longs extraits avec les conditions du point B. En général, les motifs sont groupés selon leurs évolutions de croissance de la longueur pour mettre en évidence les modifications du support, de la connexité, des couvertures et de la pureté globale. La notation M avant le motif signifie un motif maximal (qui n'a pas de sur-motif SFG).

No	Motif	SR [%]	CM	CRSM	CVT [%]	CCP [%]	PG [%]	CP (secondaire)	Fig
B ($s = 2\,;\ \sigma_{rel} = 0,5\%\,;\ \kappa = 5$)									
1	18 - 2x12_1x6	13,94	6,08	169,5	15,59	37,13	83,63	Blé(mout)	
2	18 - 2x13_1x5	16,98	6,50	220,7	30,87	75,04	86,62	Blé	
3	19 - 2x13_1x6	9,03	5,69	102,8	13,57	32,25	83,44	Blé	
4	18 - 2x14_1x4	11,99	6,05	145,1	23,74	55,09	82,78	Blé	6.38a)
5	19 - 2x14_1x5	9,86	5,94	117,1	19,82	49,93	88,71	Blé	6.38b)
6	M 20 - 2x14_1x6	3,96	5,06	40,1	7,15	15,82	77,62	Blé	6.38c)
7	18 - 2x15_1x3	5,51	5,26	58,0	6,96	17,81	90,34	Blé	6.42b)
8	19 - 2x15_1x4	3,98	5,25	41,8	6,94	17,55	90,86	Blé	
9	M 20 - 2x15_1x5	2,64	5,05	26,7	4,86	12,42	89,58	Blé	
10	18 - 1x12_2x6	11,44	6,03	138,0	20,45	74,74	50,06	Pp(Soudan)	
11	19 - 1x12_2x7	4,23	5,98	50,6	5,46	30,72	54,11	Pp(Soudan)	
12	M 20 - 1x12_2x8	1,69	5,26	17,8	2,81	15,55	52,50	Pp(Soudan)	
13	18 - 1x13_2x5	17,13	6,14	210,4	22,39	68,16	48,62	Soudan(Pp)	
14	19 - 1x13_2x6	8,11	5,64	91,5	14,84	61,52	45,18	Soudan(Pp)	
15	18 - 1x14_2x4	17,90	6,30	225,5	30,42	75,59	76,36	Maïs	6.39a)
16	19 - 1x14_2x5	11,94	5,76	137,5	14,75	23,87	66,00	Soudan(Pp)	6.39b)
17	M 20 - 1x14_2x6	4,11	5,07	41,7	7,28	37,87	51,15	Soudan(Pp)	6.39d)
18	18 - 1x15_2x3	12,32	6,09	150,1	21,79	73,14	88,75	Maïs	6.40c)
19	19 - 1x15_2x4	10,56	5,88	124,2	17.80	60.12	89.32	Maïs	6.39c)
20	18 - 1x16_2x2	5,35	5,49	58,7	12,48	42,37	89,83	Maïs(soja)	
21	19 - 1x16_2x3	5,16	5,53	57,1	12,47	42,29	91,83	Maïs	
22	M 20 - 1x16_2x4	3,71	5,22	38,7	8,77	29,41	88,69	Maïs	6.39e)
23	18 - 1x18	4,04	6,10	49,3				Eau	
24	19 - 1x19	3,35	6,41	42,9				Eau	
25	M 20 - 1x20	2,74	6,64	36,4				Eau	
26	M 18 - 1x2_2x11_1x5	4,65	5,42	50,4	5,63	15,67	97,62	Blé	
27	M 18 - 1x2_2x12_1x4	4,33	5,39	51,1	6,24	15,65	94,03	Blé(colza)	
28	M 18 - 1_2x11_1x6	6,27	5,05	63,3	7,69	21,80	99,54	Blé	
29	18 - 1_2x12_1x5	10,39	6,02	125,1	18,59	51,35	96,91	Blé pur	
30	18 - 1_2x13_1x4	8,71	5,82	101,4	17,67	47,45	95,86	Blé(colza)	
31	M 19 - 1_2x13_1x5	7,64	5,69	86,9	8,59	22,91	93,57	Blé	
A ($s = 3\,;\ \sigma_{rel} = 0,5\%\,;\ \kappa = 5,5$)									
32	M 18 - 3x15_1x3	1,19	5,58	13.3	3,95	11,10	98,80	Blé	6.42a)
33	M 18 - 1x15_3x3	4,37	5,51	48.2	13,61	47,06	91,60	Maïs(soja)	6.40a)
34	M 18 - 1x14_2_3x3	7,03	5,70	80.1	19,58	66,75	91,13	Maïs(soja)	6.40b)
35	M 18 - 1x11_2_3x6	3,54	5,92	41.9	5,34	32,87	69,05	Pp(Soudan)	
36	M 18 - 1x12_3x6	3,71	5,79	43.0	9,00	50,03	59,02	Soudan(Pp)	
37	18 - 1x18	2,62	6,94	36.4				Eau	
38	19 - 1x19	2,43	6,96	33.8				Eau	
39	M 20 - 1x20	2,20	6,95	30.6				Eau	

TAB. 6.7 – Les principaux motifs longs extraits avec les conditions des points d'opération B ($s = 2\,;\ \sigma_{rel} = 0,5\%\,;\ \kappa = 5$) et A ($s = 3\,;\ \sigma_{rel} = 0,5\%\,;\ \kappa = 5,5$) (Pp=Petit pois, mout=moutarde)

La croissance de la longueur a comme effet la décroissance du support relatif, de la connexité moyenne, (de cette manière également la diminution de la connexité relative au support relatif et de la connexité globale - un vrai coefficient de la qualité d'un motif), de la couverture globale et de la couverture de la culture principale dans la vérité terrain. À l'exception des groupes de motifs 1 - 2 - 3, 10 - 11 - 12 et 31 - 32 pour lesquels la pureté globale décroît tandis que la longueur augmente, les motifs présentent une oscillation de la valeur de cette pureté. La croissance de la pureté globale, avec la croissance de la longueur, peut être expliquée par la disparition accentuée des pixels des cultures secondaires. L'eau présente une situation spéciale : en dépit de la décroissance normale de la connexité globale avec l'augmentation de la longueur, la connexité moyenne croît pour le point d'opération B. Avec le développement en longueur du motif de l'eau, les pixels qui se perdent avec priorité sont les pixels de frontière, probablement plus faiblement liés. Une situation différente pour la connexité moyenne des motifs de l'eau apparaît dans le cas de la croissance du nombre des symboles. Pour $s = 3$ et $s = 4$, la connexité moyenne de ces motifs a des valeurs presque constates : d'environ 6,95 et respectivement 7,00.

Si on a un motif de longueur L, on étudie comme exemple les conséquences du passage vers le sur-motif adjacent de longueur $L + 1$, le motif 2x14_1x4 décrit avec $s = 2$, extrait sous les contraintes définies par les seuils $\sigma_{rel} = 0,5\%$ et $\kappa = 5$ (Figure 6.38a). Le nouveau motif s'obtient par l'apparition d'un nouveau symbole «1» ou «2» dans la description de l'évolution.

Par exemple, un nouveau symbole «1» peut apparaître dans les suivantes positions :

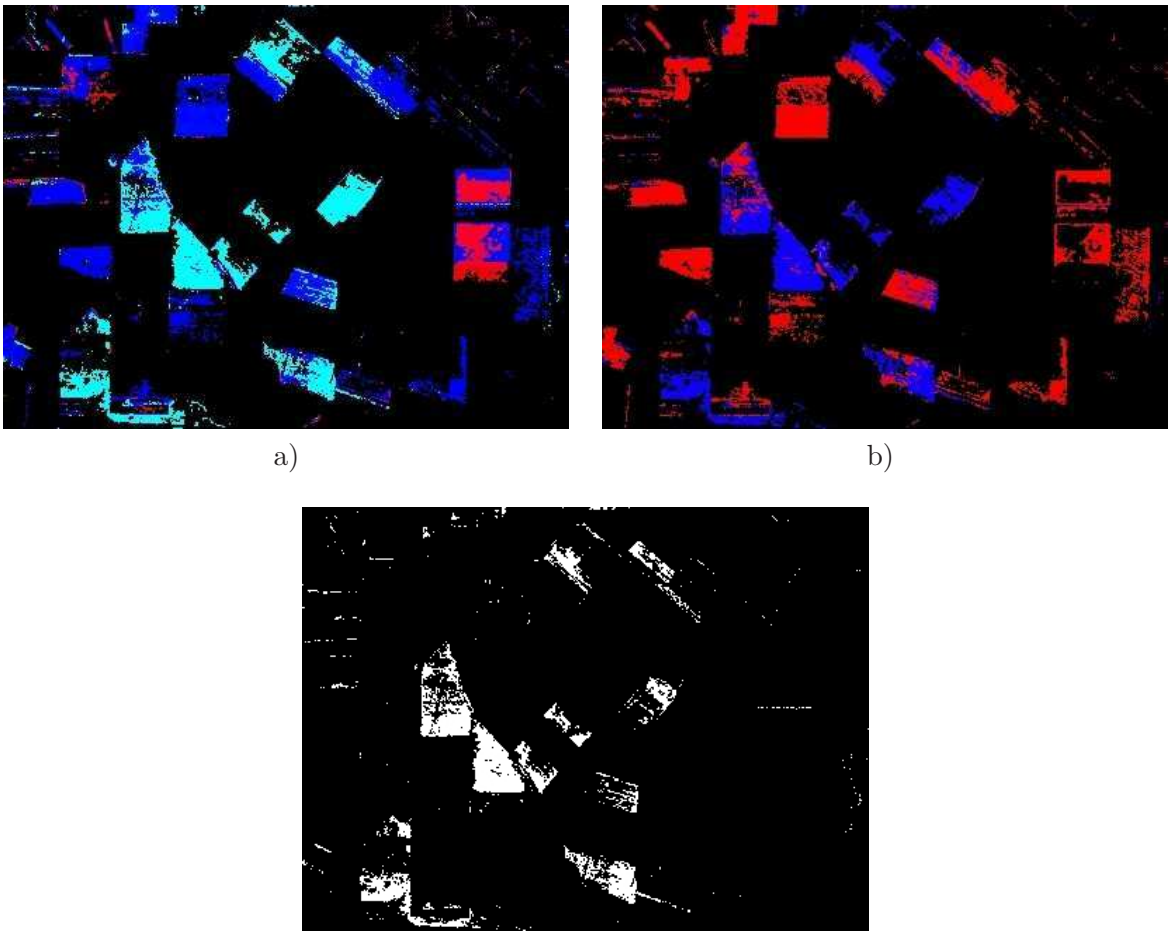
1. avant le groupe de «2» ; le motif ne respecte pas les contraintes ;
2. à l'intérieur du groupe de «2» ; le motif ne respecte pas les contraintes et l'évolution phénologique non plus ;
3. après le groupe de «2» ; en résulte le motif 2x14_1x5 valable

Un nouveau symbole «2» conduit également à trois possibilités et seul le motif 2x15_1x4 est valable. Cela explique la diminution du support, de la couverture de la vérité terrain, de la couverture de la culture principale et même de la connexité globale et de la connexité relative au support minimum dans le cas de passage à un sur-motif. Bien sûr qu'un sur-motif peut résulter des plusieurs sous-motifs adjacentes mais ces diminutions se manifestent en comparaison avec chaque sous-motif.

La variation de la connexité moyenne dépend de la spécificité thématique de la scène sous observation et dans le cas de la STIS ADAM il y a une tendance à la décroissance pour les objets agricoles. Pour un objet très connexe, le cas de l'eau, la tendance est contraire. La croissance de la longueur du motif qui représente l'eau, de 1x18 à 1x20, a comme effet les diminutions du support et de la connexité globale mais une augmentation de la connexité moyenne (voir les positions $23 - 25$ du Tableau 6.7).

La variation de la pureté globale est difficile à quantifier et dépend des conditions locales des cultures définies comme principales dans les variantes du motif. Dans le tableau on peut voir l'oscillation de la pureté globale au cours de la spécialisation du motif, respectivement de la croissance de sa longueur.

Quelques effets de la croissance de la longueur d'un motif sont visibles dans la suite des images de la Figure 6.38. Les motifs des images ont la culture principale le blé et comme cultures secondaires la moutarde et le pois chiche qui apparaissent ensemble même dans les motifs complets. Le 20-motif est maximal mais il y a des réductions du support, de différentes sortes de connexité, des couvertures de la vérité terrain et de la culture principale. Le motif reste pluriculture. En spécialisant le motif initial, des régions entières disparaissent et ceux qui restent décroissent en superficie et leurs contours sont de moins en moins nets. C'est un exemple qui justifie la considération de motifs plus courts que les motifs maximaux.



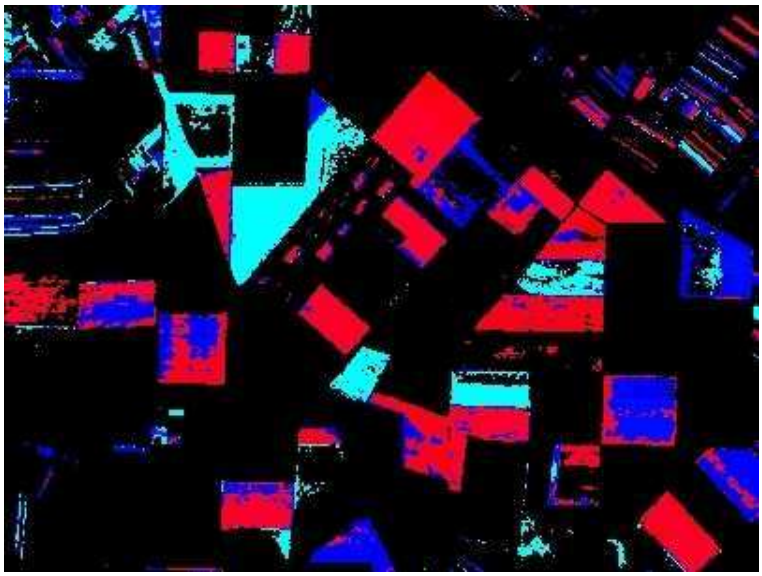
a) b)

c)

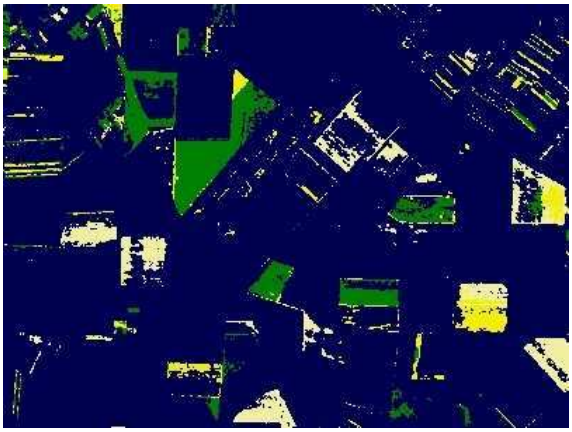
FIG. 6.38 – La spécialisation du 18-motif 2x14_1x4 : a) Le 18-motif 2x14_1x4 ($SR = 11,99\%; CM = 6,05; CRSM = 145,1; CCP = 55,09\%; PG = 82,78\%$) ; b) Le 19-motif 2x14_1x5 ($SR = 9,86\%; CM = 5,94; CRSM = 117,1; CCP = 49,93\%; PG = 88,71\%$) ; c) Le 20-motif 2x14_1x6 ($SR = 3,96\%; CM = 5,06; CRSM = 40,1; CCP = 15,82\%; PG = 77,62\%$).

Dans la Figure 6.39 est présentée l'évolution de la couverture du 18-motif 1x14_2x4 au cours de sa spécialisation vers un 20-motif. Le 18-motif initial est un exemple de la discrimination de ces variantes temporelles : maïs en rouge et en bleu, le petit pois et l'herbe du Soudan en bleu ciel. En augmentant la longueur et en conservant le groupe 1x14 (l'évolution $a \rightarrow b \rightarrow d$), le motif se «spécialise» en décrivant le petit pois et l'herbe du Soudan. Le motif complet obtenu ne discrimine pas le petit pois de l'herbe du Soudan et la couverture est faible. La conservation de l'ensemble «2x4» conduit à une culture de maïs suffisamment pure (l'évolution $a \rightarrow c \rightarrow e$). Dans la Figure 6.39c) est observé un saut en pureté, le maïs reste seul parce que son cycle est translaté après les cycles de ses compagnons dans le motif antérieur (leurs cycles ont beaucoup de «2» dans la période de temps d'observation). Dans l'évolution $c \rightarrow e$, la couverture baisse, des régions utiles disparaissent, mais la pureté globale est stable, les motifs étant de candidats suffisament bons pour un éventuel clustering.

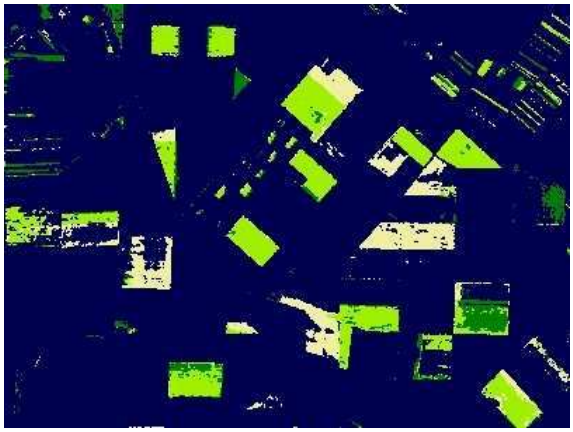
Le compromis fait par le choix d'un nombre réduit de symboles pour la valeur des pixels affecte naturellement la précision de description des évolutions phénologiques par les MSFG extraits. Néanmoins, ce choix tire profit de la couverture supérieure de la scène et n'affecte pas, d'une manière significative, la précision de la correspondance entre les motifs extraits et les différents types de cultures agricoles. Une discussion sur le thème des influences induites par la



a)

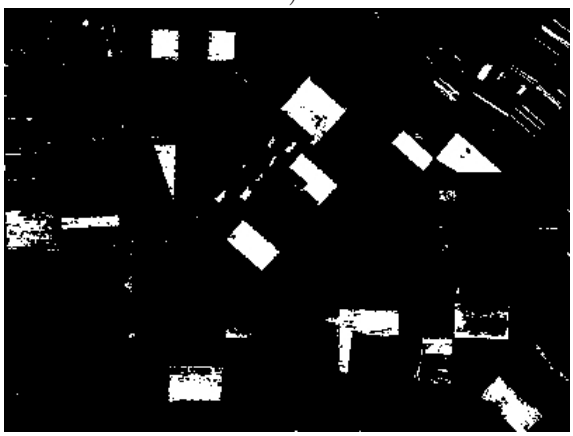


b)



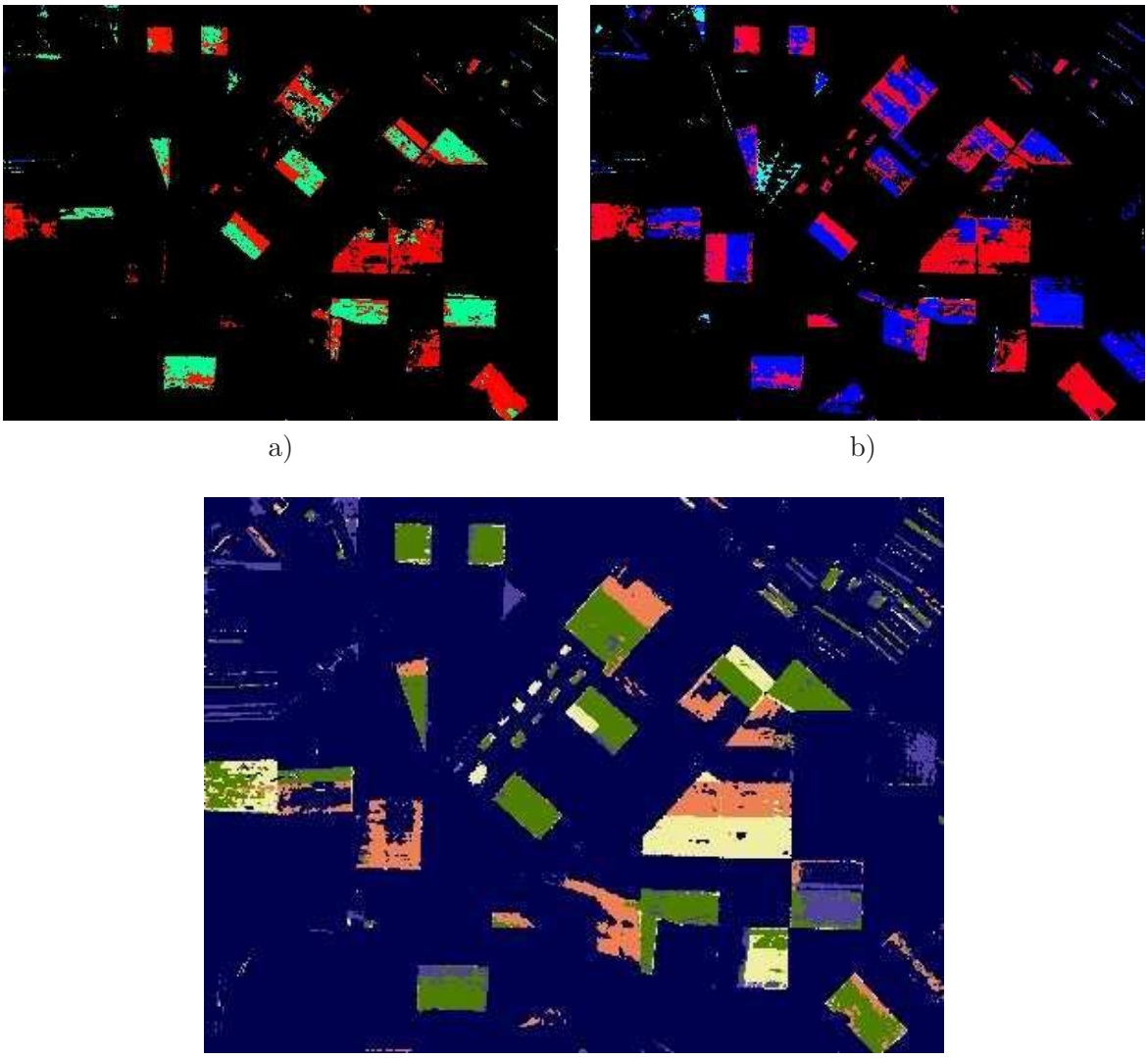
c)

d)



e)

FIG. 6.39 – La spécialisation du 18-motif 1x14_2x4 : a) Le 18-motif 1x14_2x4 ($SR = 17,90\%; CM = 6,30; CRSM = 225,5; CVT = 21,79\%; CCP = 73,14\%; PG = 88,75\%$) ; b) Le 19-motif 1x14_2x5 ($SR = 11,94\%; CM = 5,76; CRSM = 137,5; CVT = 14,75\%; CPP = 23,87\%; PG = 66,00\%$) ; c) Le 19-motif 1x15_2x4 ($SR = 10,56\%; CM = 5,88; CRSM = 124,2; CVT = 17.80\%; CCP = 60.12\%; PG = 89.32\%$) ; d) Le 20-motif 1x14_2x6 ($SR = 4,11\%; CM = 5,07; CRSM = 41,7; CVT = 1,19\%; CPP = 2,04\%; PG = 60,03\%$) ; e) Le 20-motif 1x16_2x4 ($SR = 3,71\%; CM = 5,22; CRSM = 38,7; CVT = 8,77\%; CCP = 29,41\%; PG = 88.69\%$).



a) b)

c)

FIG. 6.40 – Correspondance entre des motifs semblables extraits avec $s = 3$ et $s = 2$: a) Le 18-motif 1x15_3x3 ($s = 3; SR = 4,37\%; CM = 6,51; CRSM = 48,2; CVT = 11,93\%; CCP = 40,74\%; PG = 90,33\%$) ; b) Le 18-motif 1x14_2_3x3 ($s = 3; SR = 7,03\%; CM = 5,70; CRSM = 83,2; CVT = 19,58\%; CCP = 66,75\%; PG = 91,13\%$) ; c) Le 18-motif 1x15_2x3 ($s = 2; SR = 12,32\%; CM = 6,09; CRSM = 150,1; CVT = 21,79\%; CCP = 73,14\%; PG = 88,75\%$).

variation du nombre de symboles est utile et pertinente.

Pour une correspondance correcte des motifs sont choisis, pour $s = 3$, des motifs contenant seulement les symboles «1» et «3» qui se transforment, pour $s = 2$, dans les symboles «1» et «2». Par exemple, le 18-motif 1x15_3x3 ($s = 3$) et son correspondant le 18-motif 1x15_2x3 ($s = 2$) sont présentés dans la Figure 6.40 a) et c) respectivement.

Les motifs décrivent des comportements phénologiques semblables et la culture principale est le maïs. Si le nombre de symboles décroît, de $s = 3$ à $s = 2$, on gagne considérablement en support, en connexité globale et relative au support minimum et en tous les types de couvertures mais on perd en connexité moyenne et en pureté globale. Pour les applications, on peut dire que le gain en couverture surclasse la petite perte en pureté.

En particulier, le motif 1x15_2x3 a le meilleur produit $CCP \times PG$ qui peut constituer un bon facteur de qualité pour la couverture de culture dans la vérité terrain. Les bons résultats du 18-

motif 1x15_2x3 peuvent être expliqués par la contribution du motif 1x14_2_3x3 (Figure 6.40b) qui pour $s = 3$ a le meilleur facteur de qualité décrit au-dessus. Ce motif permet la discrimination temporelle de la culture de l'herbe du Soudan (en bleu ciel) de la culture principale qui est le maïs (en rouge et bleu). Le support relatif est le meilleur pour les motifs extraits dans les conditions du point d'opération A.

Si on fait la comparaison entre le 1x14_2_3x3 ($s = 3$) et le 18-motif 1x15_2x3 ($s = 2$) les résultats sont presque les mêmes que pour la comparaison antérieure. La seule différence est qu'ici la connexité moyenne est meilleure pour le cas $s = 2$.

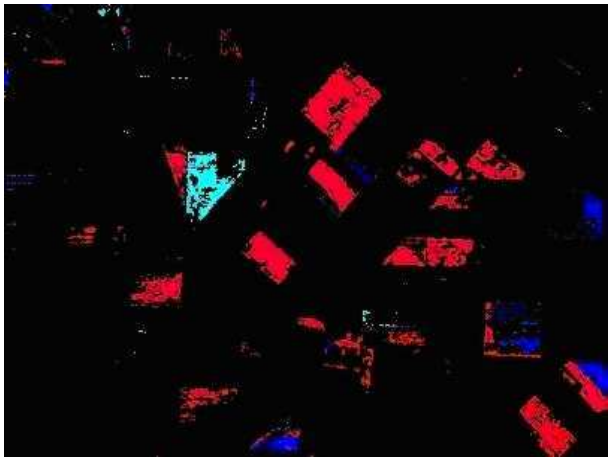


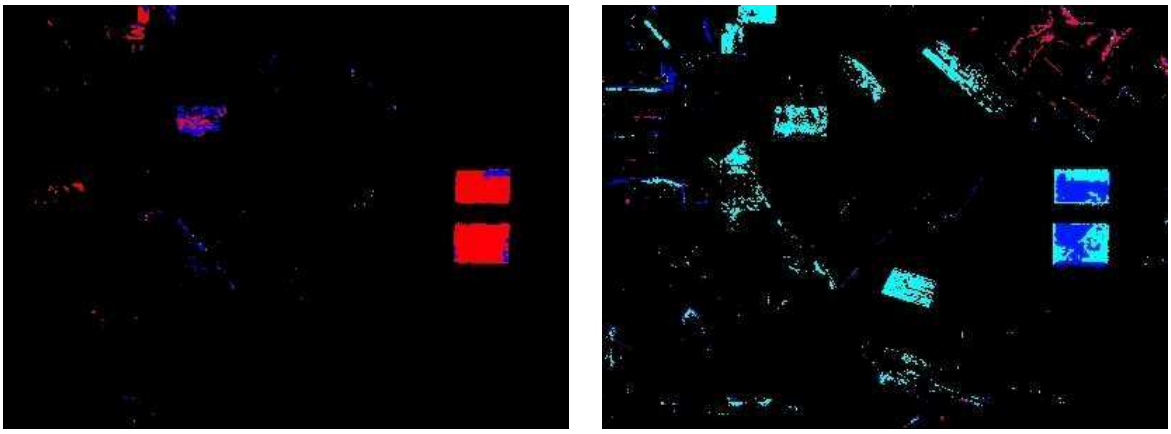
FIG. 6.41 – La localisation du 18-motif 1x14_3_4x3 ($s = 4$)

Si on fait l'extraction avec le nombre de symboles $s = 4$, pour obtenir un motif semblable, par exemple le 18-motif 1x14_3_4x3 (Figure 6.41), il est nécessaire de baisser le seuil de connexité moyenne. Pour la valeur du seuil de la connexité moyenne $\kappa = 5$, avec des données IVDN, on obtient seulement 7 motifs, en majorité décrivant le blé.

Le motif correspondant pour $s = 4$ est le motif le plus peuplé de l'ensemble de motifs extraits avec les conditions $\sigma_{rel} = 0,5\%$, $\kappa = 4$. Le support relatif est $2,08\%$ et la connexité moyenne a la valeur 4,86. Il offre une image bruitée et la possibilité de la discrimination temporelle de l'herbe de Soudan (en bleu ciel) en comparaison avec le maïs qui est représenté en rouge et bleu. La croissance du nombre de symboles, de la valeur 3 à 4, conduit dans ce cas à la diminution en même temps du support et de la connexité moyenne, fait qui a été mis en évidence également dans l'étude de la variation du nombre de motifs.

Un exemple avec la culture du blé est présenté dans la Figure 6.42. Pour le nombre de symboles $s = 3$, le 18-motif 3x15_1x3 (Figure 6.42a) a la meilleure pureté globale et met en évidence avec une bonne précision une certaine sorte de blé. Son correspondant pour $s = 2$, le 18-motif 2x15_1x3 (Figure 6.42b) a un support et des couvertures meilleures mais la connexité moyenne et la pureté globale sont diminuées parce que des nouveaux pixels sont couverts et n'appartiennent pas, en totalité, à la même culture.

Une autre comparaison, entre les motifs décrivant l'eau, montre le «pouvoir» discriminatoire élevé d'un nombre de symboles supérieur. L'eau de la rivière Mostiştea est l'objet le plus compact de la scène. Dans notre configuration, les valeurs des pixels correspondants sont presque nulles. En augmentant le nombre de symboles, le support diminue ; pour le 18-motif 1x18, le support relatif est $4,04\%$ ($s = 2$) ; $2,62\%$ ($s = 3$) et $2,47\%$ ($s = 4$). L'écart des valeurs pour le symbole «1» diminue avec la croissance du nombre de symboles et les pixels décrivant l'eau sont accompagnés de moins en moins par d'autres pixels. Le poids des pixels «d'eau» augmente



a) b)

FIG. 6.42 – Comparaison entre : a) le 18-motif 3x15_1x3 ($s = 3; SR = 1,19\%; CM = 5,58; CRSM = 13,4; CVT = 3,95\%; CCP = 11,10\%; PG = 98,80\%$) et b) le 18-motif 2x15_1x3 ($s = 2; SR = 5,51\%; CM = 5,26; CRSM = 58,00; CVT = 6,96\%; CCP = 17,81\%; PG = 90,34\%$).

et la connexité moyenne s'amplifie, 6,1 ($s = 2$) ; 6,94 ($s = 3$) et 6,99 ($s = 4$).

En conclusion la croissance de la longueur et du nombre de symboles d'un motif a comme résultat la diminution du support, de la CG et CRSM et des couverture dans VT, pendant que la CM et la PG peuvent osciller.

L'utilisateur doit donc établir le compromis entre les variations opposées de la couverture et de la pureté globale.

Chapitre 7

Données Radar : les STIS du projet EFIDIR

Sommaire

La généricité du concept d'extraction de MSFG dans des STIS permet son utilisation pour des différents types de données. Dans ce chapitre la méthode est appliquée pour des données radar [120, 119, 126, 116].

Le RAdio Detection And Ranging (RADAR) est un système actif qui se fonde sur la propagation électromagnétique : une onde émise par une source se réfléchit sur des cibles, et l'analyse du signal reçu permet de détecter et de localiser ces cibles. À chaque pixel de l'image radar, on associe une valeur complexe issue du signal reçu après l'émission et la rétrodiffusion sur une surface. En supposant que la vitesse de propagation est constante, toute mesure de temps (par exemple le délai entre émission et réception) peut se traduire en mesure de distance. Le radar utilise des fréquences comprises entre 0,3 et 300 GHz correspondant à des micro-ondes de 1m à 1mm [103]. Il permet d'imager la terre sans soucis d'éclairement solaire (jour et nuit), il n'est pas autant affecté par la couverture nuageuse ou la brume que les capteurs optiques. En effet, les ondes radar traversent les perturbations atmosphériques. Néanmoins, le délai de propagation peut être modifié par les conditions météorologiques. Ces ondes peuvent caractériser les objets : que ce soit leur position horizontale, leur altitude, leur vitesse et parfois leur forme.

Un Radar à Synthèse d'Ouverture (en anglais Synthetic Aperture Radar, SAR) (RSO) est un système radar générant des images de télédétection à haute résolution. Pour créer une image, sont utilisées l'amplitude et la phase du signal reçu de plusieurs impulsions successives pour simuler une plus grande ouverture (ou dimension) d'antenne, d'où le terme «synthèse d'ouverture».

Une des principales différences géométriques entre l'imagerie optique et l'imagerie radar est que cette dernière ne met en œuvre que des distances et non des angles [153]. La spécificité de la géométrie des images radar induit des distorsions liées à l'échantillonnage en distance. En effet, les distances entre les points du sol ne sont pas conservées lors de la formation des images radar. Celles-ci contiennent des dilatations, des compressions, des recouvrements et des zones d'ombre selon l'orientation par rapport à l'angle de visée du radar [113, 127, 128]. Les distorsions sont d'autant plus importantes que le terrain présente de fortes dénivellations. Ainsi dans une zone montagneuse, les versants orientés vers le radar se rétrécissent voire se superposent et les versants opposés s'allongent [184, 183].

Le signal RSO contient des informations d'amplitude et de phase. Deux techniques différentes se sont développées afin d'exploiter l'information de phase. La première technique utilisant l'information de phase est l'interférométrie. Elle est fondée sur la différence de phase entre deux signaux radar complexes obtenus en imageant deux fois la même zone. En supposant que le terme de phase propre (dépendant de la cible) est le même pour les deux images, la différence de phase devient proportionnelle à la variation de la distance aller-retour radar-cible. La deuxième technique étudie la signature du terrain. Pour cela on forme la différence de phase entre deux images acquises simultanément avec des configurations de polarisation différentes. La polarisation est définie comme l'orientation du vecteur électrique d'une onde électromagnétique. Les antennes d'un système radar peuvent être configurées de façon à émettre et à capter un rayonnement électromagnétique polarisé horizontalement ou verticalement. Lorsque l'onde radar atteint une surface est en est diffusée, la polarisation peut être modifiée, en fonction des propriétés de la surface. Cette technique appelée polarimétrie permet de discriminer, par exemple, certains types de végétations lorsqu'ils présentent de fortes asymétries géométriques.

Le projet EFIDIR [79] a pour but de développer une plateforme ouverte d'archivage et de traitements adaptée d'une part aux spécificités des données RSO et d'autre part aux grandes séries temporelles exploitées pour les mesures de déplacement. Le projet s'appuie sur des bases de données liées à plusieurs thématiques telles que : les mouvements de faible amplitude (petits et lents, $\sim$ quelques mm/an) mais de grande extension spatiale (grande longueur d'onde spatiale, $\sim$ 100 km) liés au remplissage de grands barrages, les mouvements localisés, de plus grande

amplitude, de surface des glaciers et les mouvements volcaniques. Parmi les STIS étudiées dans le cadre du projet figurent les zones du lac Mead et de Chamonix Mont-Blanc.

7.1 La STIS du lac Mead - Interférométrie radar

Les images RSO, complexes, suscitent un grand intérêt pour la double information qu'elles apportent : l'amplitude et la phase.

Les images d'amplitude représentent la réponse du terrain à l'onde émise par le radar, aux atténuations de transmission près. L'amplitude appelée aussi radiométrie du pixel est fonction de l'intéraction onde-matière sur la surface imagée correspondante. Elle dépend de deux ensembles de paramètres : les paramètres propres au radar (longueur d'onde, polarisation, angle d'incidence) et les paramètres liés à la nature du sol (réflectance, humidité, rugosité de la surface par rapport à la longueur d'onde, inclinaison du sol, propriétés diélectriques).

La phase comporte une composante géométrique utile liée à la propagation aller-retour de l'onde électromagnétique et des composantes liées à la trajectoire orbitale, aux conditions atmosphériques et instrumentales et au mécanisme de rétrodiffusion de la cible. Cette dernière composante nommée phase propre dépend des paramètres tels que la pénétration des ondes, la constante diélectrique, la répartition des réflecteurs élémentaires.

La phase étant connue en valeur principale (modulo 2π), la mesure de distance de radar-cible est accessible modulo $\lambda/2$. Malgré son ambiguïté, cette mesure présente un grand intérêt du fait de sa précision de l'ordre d'une fraction de longueur d'onde. À partir de la différence de phase entre deux images se distinguant par leurs dates d'acquisition ou par leurs configurations spécifiques [182], on peut établir une nouvelle image appelée interférogramme qui représente des franges prenant des valeurs allant de 0 à 2π.

En accédant à la composante géométrique par différence de phases entre deux acquisitions, l'interférométrie radar satellitaire multi-passes fournit une mesure jusqu'ici inaccessible en de nombreux sites : la mesure du déplacement au sol entre deux dates avec un pas d'une dizaine de mètres et une précision de l'ordre de la longueur d'onde (pour la bande C entre 3,75 et 7,5 cm [200]).

Si l'acquisition se fait sous le même angle mais à des moments décalés, on obtient un interférogramme différentiel. L'image correspondante est caractéristique des changements tridimensionnels qui sont intervenus entre les acquisitions. Les interférogrammes différentiels permettent d'étudier les modifications du relief causées par un tremblement de terre, une éruption volcanique, un glissement de terrain, une dérive glacière, etc. Dans le cas de l'observation des glaciers (étude du déplacement de la glace [209, 213, 208] et, plus généralement, pour la détection de changement, les interférogrammes différentiels sont obtenus en réalisant la différence de deux images acquises à quelques jours d'intervalle. Des intervalles de temps supérieurs, de l'ordre du mois ou de l'année, peuvent être intéressants pour l'étude des phénomènes sismiques, volcaniques ou d'affaissement du sol.

Le coefficient de corrélation entre les deux images rend compte de la similarité des mécanismes de rétrodiffusion des deux acquisitions. Ce paramètre est nommé la cohérence. Elle est très sensible au bruit présent dans l'interférogramme et peut être perçue à ce titre comme un indicateur de fiabilité de la différence de phase. Ainsi, les zones de faible cohérence, peu pertinentes, peuvent être mises de côté lors du déroulement de phase. Ce processus de déroulement de phase consiste à lever l'ambiguïté de la phase interférométrique et à trouver une estimation de la phase absolue pour chaque cible (pixel) à partir de la phase mesurée (modulo 2π).

7.1.1 Données, pré-traitements des données et phénoménologie de la scène

Les STIS d'interférogrammes différentiels sont des ensembles de données difficiles, car elles représentent de gros volumes de données, et, comme les acquisitions, elles sont influencées par les conditions atmosphériques. Dans ce chapitre, sont considérées des images RSO couvrant la région du lac Mead (Nevada, États-Unis). Elles ont été acquises par les satellites européens de télédétection, en anglais ERS 1 et 2, et par ENVironmental SATellite (ENVISAT) et mises à disposition par l'intermédiaire du projet EFIDIR.

Le satellite ERS-1 a été lancé en 1991, et ERS-2 en 1995 (dans le même plan orbital que ERS-1) par l'Agence Spatiale Européenne (ESA). Ils sont placés sur une orbite quasi-circulaire inclinée à 98,5° et d'altitude moyenne de 785 km [2]. Ces satellites disposent d'instruments permettant d'étudier les phénomènes glaciaires, la météorologie, ainsi que les phénomènes accessibles aux techniques de télédétection, notamment par l'utilisation du radar à synthèse d'ouverture (RSO). Les satellites ERS sont dotés de 6 instruments, dont un capteur RSO qui opère dans la bande C ($4 - 8$ GHz), ayant une polarisation VV (transmission verticale du signal, réception verticale du signal) et une longueur d'onde de 5,66 cm.

ENVISAT est un satellite mis en orbite en mars 2002 par Ariane 5 depuis le Centre spatial guyanais de Kourou, en Guyane française [3]. Ce programme de l'ESA a pour objectif d'assurer la continuité des missions ERS, tout en apportant des observations de paramètres additionnels (observations dans différentes polarisations ou combinaisons de polarisations, différents angles d'incidence et différentes résolutions spatiales, le tout en bande C) afin de contribuer efficacement à l'étude de l'environnement. ENVISAT évolue à une altitude moyenne de 800 km sur une orbite quasi-circulaire, inclinée de 98,6° par rapport au plan équatorial, ce qui lui confère l'héliosynchronisme. Il embarque dix instruments scientifiques, complétés par le système de positionnement DORIS.

Le lac Mead est le plus grand réservoir artificiel d'eau des États-Unis et il est situé dans le désert de Mojave entre le Nevada et l'Arizona, à 48 km sud-est de Las Vegas (Figure 7.1). Il a été créé en 1935 après la construction du barrage Hoover. Il a une surface d'environ 640 km^2 et contient environ 35 km^3 d'eau. Le lac est formé de plusieurs bassins alimentés par le fleuve Colorado et ses affluents. Les bassins Boulder et Virgin prédominent et représentent environ 60% du volume total d'eau. Le lac a une altitude d'environ 350 m et est bordé par des montagnes orientées nord-sud. L'altitude de la zone, 700 m en moyenne, augmente vers l'est jusqu'à 1500 m sur le plateau du Colorado.

Le niveau d'eau a subi des fluctuations inter-annuelles d'environ 20 m pour la période 1992-2009. La surface du sol autour du lac est touchée par un mouvement d'affaissement/soulèvement qui est en corrélation avec les fluctuations du niveau d'eau. À une échelle de 50 km, la surface est affectée par une subsidence lorsque le niveau d'eau augmente et inversement lorsque le niveau d'eau diminue [45]. D'autre part, localement, quelques zones se soulèvent lorsque le niveau d'eau augmente. En effet, la pression accrue de l'eau provoque la dilatation des sols.

La différence de phase interférométrique entre deux images radar (image esclave moins image maîtresse) est calculée pour mesurer la déformation du sol. Toutefois, elle ne contient pas seulement l'effet du mouvement du sol dans la ligne de visée du radar, mais aussi des erreurs résiduelles orbitales, des délais dans l'atmosphère, et du bruit. La plupart des interférogrammes montre des artefacts atmosphériques forts, d'amplitude plus grande que le mouvement prévu du sol, de deux types : (1) délais de phase corrélés avec l'élévation résultant de la variation temporelle de la stratification de vapeur d'eau dans la troposphère et (2) délais de phase de formes variables dans le temps et l'espace, sous la forme de petites ondulations, de taches floues, de grandes taches ou de fronts.

FIG. 7.1 – La zone du lac Mead (carte des sites de collecte de données `http://www.wakelv.com/main/usgs/`).

Pour caractériser la déformation de la croûte terrestre associée aux fluctuations du niveau du lac, un sous-ensemble de 20 interférogrammes obtenus à partir d'images acquises entre 1996 et 2008 sont sélectionnés. Une série d'interférogrammes de faible ligne de base est utilisée pour récupérer l'évolution temporelle du changement de phase, pour chaque pixel d'une scène RSO. Chaque interférogramme donne la différence de phase interférométrique de sa date d'acquisition par rapport à la date maîtresse 08/10/1995.

Des corrections des erreurs d'orbite, qui se traduisent comme des courbures de la phase avec un axe de symétrie parallèle à l'azimut, sont effectuées pour chaque interférogramme. De plus, les délais de phase troposphérique sont estimés et corrigés pour chaque interférogramme par l'analyse de la corrélation phase - élévation [45]. La corrélation entre le changement en distance et l'élévation est due à la variation entre deux acquisitions RSO du contenu moyen en vapeur d'eau dans l'atmosphère la plus basse. Les perturbations atmosphériques qui restent sont spatialement aléatoires pour des dates différentes d'acquisition, alors que les modèles de déformation doivent présenter une certaine corrélation spatiale avec le temps.

Les perturbations atmosphériques de l'image maîtresse et de l'image esclave sont présentes dans chaque interférogramme. Ces perturbations atmosphériques introduisent un délai de l'onde électromagnétique dans l'atmosphère car la vitesse de propagation est légèrement inférieure à celle de la vitesse de la lumière. L'indice de réfraction varie avec la pression, la température et l'humidité (vapeurs d'eau).

Le but de cette application est de vérifier la possibilité d'extraire les déformations de la croûte terrestre en écartant les influences des perturbations atmosphériques en utilisant les MSFG.

Entre 1996 et 1998, le niveau d'eau du lac a augmenté, tandis qu'entre 2000 et 2008 il a diminué. Les images analysées (759×716 pixels, 130×130 m de résolution) contiennent des délais de phase dus aux effets atmosphériques et de déformation pour une superficie d'environ 100×100 km^2. La Figure 7.2 présente une telle image. Le délai montré comprend les effets atmosphériques

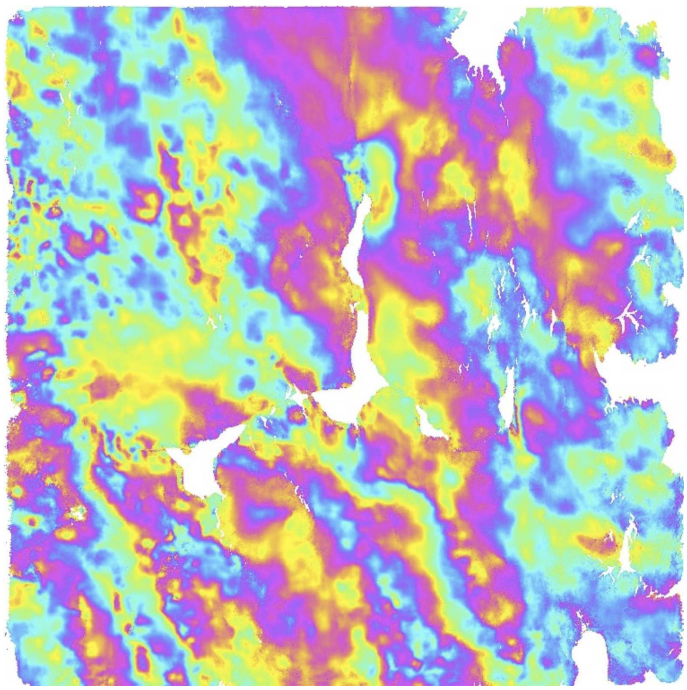


FIG. 7.2 – Délai de phase interférométrique de 08/08/1996, relativement à la date maîtresse 08/10/1995, affiché en géométrie radar.

dominants ainsi que la déformation de la surface entre 08/10/1995 et 08/08/1996. Un cycle de couleur (rouge / jaune / vert / bleu / violet) correspond à une augmentation de la distance entre le satellite et la surface de la Terre de 1,8 cm.

Indépendamment, un soulèvement continu (jusqu'à 2 cm) a été observé aussi, près de Las Vegas, au cours de la période 1992-2002. Ce mouvement du sol résulte de la déformation du système aquifère [11].

Dans la région du lac Mead, en raison des conditions arides et de l'absence de végétation, la cohérence est préservée sur une grande partie de la surface à travers des périodes étendues de temps. En raison de la forte cohérence, la phase peut être spatialement déroulée sur environ 80% de la scène radar en moyenne. Les zones blanches centrales correspondent au lac Mead sur lequel aucun délai de phase ne peut être mesuré. Les autres zones blanches correspondent au canyon de Colorado, à des sommets de montagnes et à quelques plaines pour lesquels la cohérence réduite n'a pas permis le déroulement de la phase.

Les résultats présentés ici sont obtenus à partir d'images avec des valeurs de pixels, c'est-à-dire les valeurs de différence de phase, quantifiées en 3 intervalles. Le premier intervalle (symbole «1») désigne de fortes valeurs négatives, le second (symbole «2») contient les valeurs positives et négatives réduites (proches de zéro), et le troisième (symbole «3») correspond à de fortes valeurs positives de la différence de phase. Une forte valeur positive correspondant à une augmentation de la distance terre-satellite est interprétée comme une subsidence, tandis qu'une valeur négative forte concerne un soulèvement. La procédure de quantification est celle utilisée pour la STIS ADAM (des intervalles non superposés établis par équirépartition, voir l'annexe A et [117, 125, 120, 119, 121, 122]).

7.1.2 Résultats quantitatifs

Pour mettre en évidence les déformations de la croûte terrestre comme conséquences des variations du niveau du lac Mead, l'utilisateur valide une discrétisation simple en trois intervalles [119]. De cette manière, les petites variations, positives ou négatives, peuvent être incluses dans une seule classe, une vraie classe tampon. En effet, l'équirepartition des valeurs en trois classes a comme résultat une largeur étroite de l'intervalle moyen qui collecte ainsi les très petites variations autour de la valeur nulle. De plus, l'utilisateur recommande des seuils de support et de connexité élevés, tenant compte de l'étendue du phenomène observé et de la structure rigide de la croûte terrestre dans la scène en surveillance.

Dans le Tableau 7.1 sont présentées les variations du nombre de MSFG extraits en fonction des seuils de support relatif, σ_{rel} et de connexité moyenne, κ.

σ_{rel} [%]	N_{MSF} ($\kappa = 0$)	N_{MSFG} ($\kappa = 5$)	N_{MSFG} ($\kappa = 6$)	N_{MSFG} ($\kappa = 7$)
0.45	693 847	173 960	10 867	632
0.92	362 169	137 139	10 252	632
1.84	176 807	105 686	10 172	632

TAB. 7.1 – L'extraction des motifs séquentiels de la base de données de la STIS du lac Mead.

Ces valeurs sont plus grandes que celles obtenues dans le cas de la STIS ADAM. Dans ce qui suit on voit que ce fait conduit à des longueurs maximales de motifs plus réduites.

Pour la STIS considérée de 20 images de 543 444 pixels chacune, pour $s = 3$ le nombre de motifs possibles, N_{MP}, est $5,23 \cdot 10^9$, le nombre de motifs séquentiels, N_{MS}, est $6,159 \cdot 10^7$ et la densité en motifs séquentiels a une valeur assez grande de $1,17\%$.

Si la notion de densité, introduite par la relation 6.2, est réutilisée pour les MSF et MSFG on peut définir la densité de motifs fréquents, ρ_{SF}, et la densité de motifs fréquents groupés, ρ_{SFG}.

$\rho_{SF} = N_{MSF}/N_{MS}$

$\rho_{SFG} = N_{MSFG}/N_{MSF}$

Les valeurs de ces densités pour les bases de données ADAM et lac Mead sont présentées dans le Tableau 7.2.

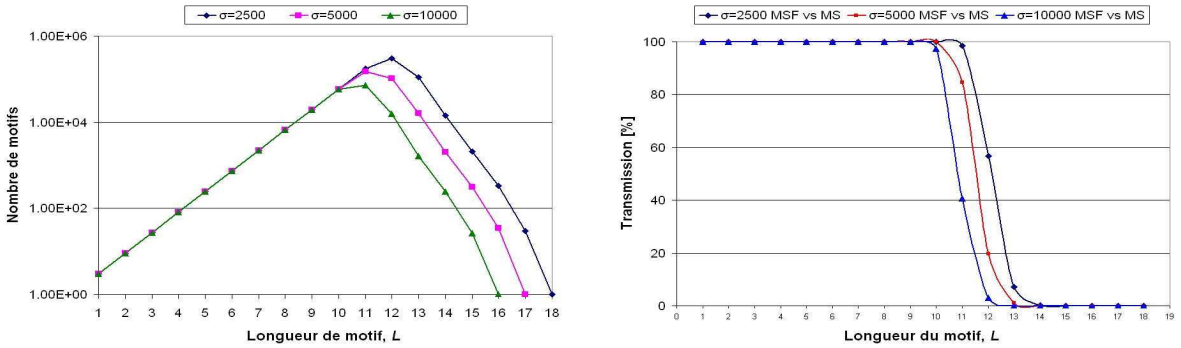
	N_{MP}	N_{MS}	ρ_S[%]	N_{MSF}	ρ_{SF}[%]	N_{MSFG}	ρ_{SFG}[%]
ADAM	$5,2 \cdot 10^9$	$10,4 \cdot 10^6$	0,20	23 038	0,22	474	2,06
MEAD	$5,2 \cdot 10^9$	$61,6 \cdot 10^6$	1,17	176 807	0,29	10 172	6,06

TAB. 7.2 – Comparaison entre les STIS ADAM et lac Mead ($s = 3; \sigma = 10.000; \kappa = \mu = 6$).

Le nombre de motifs et la densité sont plus grands pour la STIS du lac Mead. Par conséquent, les supports de ces motifs doivent être plus petits. Dans la sous-section 7.1.3, le support maximum pour les cinq motifs maximaux extraits avec les paramètres du Tableau 7.2 atteint à peine le double du support minimum. Un support réduit d'un certain motif augmente la probabilité d'élagage de ses sur-motifs. De cette manière, une densité grande conduit à des motifs de longueur maximale courte. La comparaison des fonctions de transmission des deux STIS montre des pentes de coupures vers des longueurs plus grandes pour la STIS ADAM, fait qui permet l'obtention des MSFG plus longs. La source de ces motifs longs est l'organisation thématique produite par l'intervention humaine et la présence des grandes cultures agricoles qui assure des supports élevés. La longueur maximale des motifs, obtenue pour les paramètres $s = 3$ et $\kappa = \mu = 6$, est plus petite que dans le cas de la STIS ADAM. Les valeurs sont de 15 événements pour un seuil $\sigma = 10000$ et 17 pour $\sigma = 2500$.

Le grand nombre de motifs très connexes et sa relative indépendance du seuil de support sont une conséquence de la constitution rigide du sol dans la scène (zone montagneuse). Pour des degrés élevés de connexité ($\kappa \geq 6$), la variation du nombre de motifs avec le support est très faible en confirmant l'hypothèse de départ de l'utilisateur.

Si la distribution des motifs séquentiels fréquents suivant leurs longueurs est représentée, les graphiques de la Figure 7.3a) sont obtenus.



a) b)

FIG. 7.3 – a) La distribution en longueurs des MSF de la STIS du lac Mead en fonction du seuil de support ($s = 3$) et b) Les dépendances des fonctions de transmission équivalente MS-MSF de la STIS du lac Mead suivant la longueur et le seuil de support pour $s = 3$.

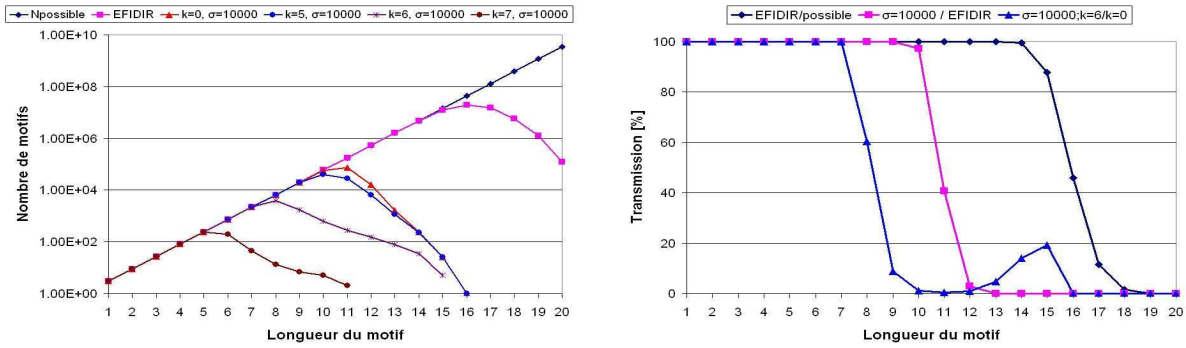
Les dépendances sont ressemblantes avec celles de la STIS ADAM, et la Figure 7.3b) confirme cette affirmation. Pour une augmentation du seuil de support, les maximums des distributions des MSF descendent et se déplacent vers les longueurs petites, une conséquence de la translation dans le même sens de la pente de coupure des fonctions de transfert équivalent. Il en est de même pour les MSFG.

La Figure 7.4a) démontre la diminution de la longueur avec l'augmentation du seuil de support.

Une autre caractéristique importante de cette STIS est le degré élevé de connexité implicite des motifs longs extraits. Ce fait est démontré par la différence insignifiante entre les distributions des plus longs MSF ($\kappa = 0$) et MSFG extraits avec un seuil assez élevé, $\kappa = 5$.

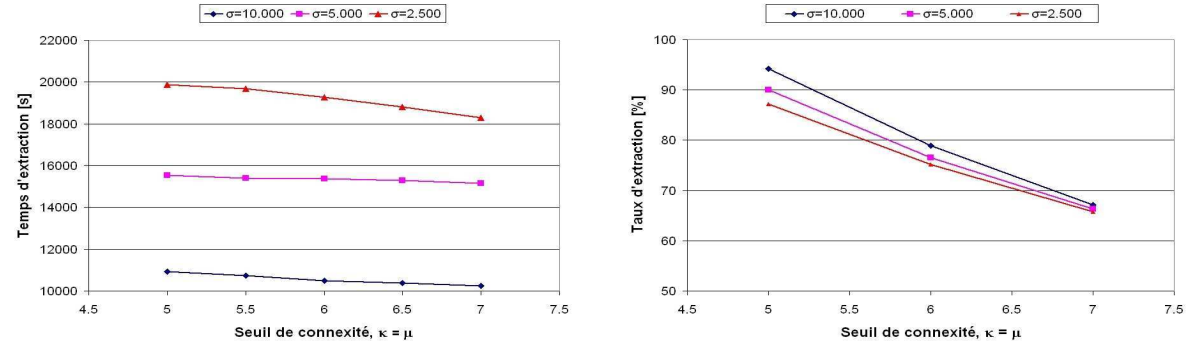
Dans la Figure 7.4b), il est possible de voir la réduction successive du nombre de motifs si les seuils de support et de connexité moyenne sont appliqués. Pour le cas agrée par l'utilisateur, un seuil de connexité élevé $\kappa = 6$, la Figure 7.4b) montre la translation de la pente de coupure des fonctions des transmission équivalente et, en plus, la présence d'une "fenêtre" dans la zone des longueurs élevées. C'est une preuve supplémentaire pour la connexité naturelle des motifs de cette STIS. Le sol a une constitution rocheuse et la connexité est implicite.

L'extraction des motifs a été réalisée avec la méthode de la relaxation de la contrainte de CM avec la contrainte de CRSM ($\kappa = \mu$). Le temps d'extraction a une petite diminution avec la croissance du seuil de connexité $\kappa = \mu$ (Figure 7.5a) et une diminution évidente avec l'augmentation du seuil de support. De point de vue du temps d'extraction, la méthode CRSM + CM ($\kappa = \mu$) donne des valeurs comprises entre celles obtenues avec les méthodes CRSM (seuil μ) et CM (seuil κ). La position relative de ces temps pour CRSM + CM ($\kappa = \mu$) dépend du nombre de motifs visités, plus exactement du taux d'extraction. La Figure 7.5b) présente la situation dans notre extraction, à savoir des grandes valeurs pour le taux, fait qui conduit à des temps placés vers l'extrémité supérieure de l'intervalle mentionnée, près des résultats extraits



a) b)

FIG. 7.4 – a) Les distributions par longueur des MSFG extraits de la STIS du lac Mead en fonction de seuils de support et de connexité moyenne, pour $s = 3$ et b) Les fonctions de transmission équivalente pour les motifs extraits de la STIS du lac Mead, pour $s = 3$.



a) b)

FIG. 7.5 – Les temps et taux d'extraction pour le paramètre $s = 3$ a) Le temps d'extraction en fonction du seuil de connexité et du seuil de support et b) Le taux d'extraction en fonction du seuil de connexité et du seuil de support.

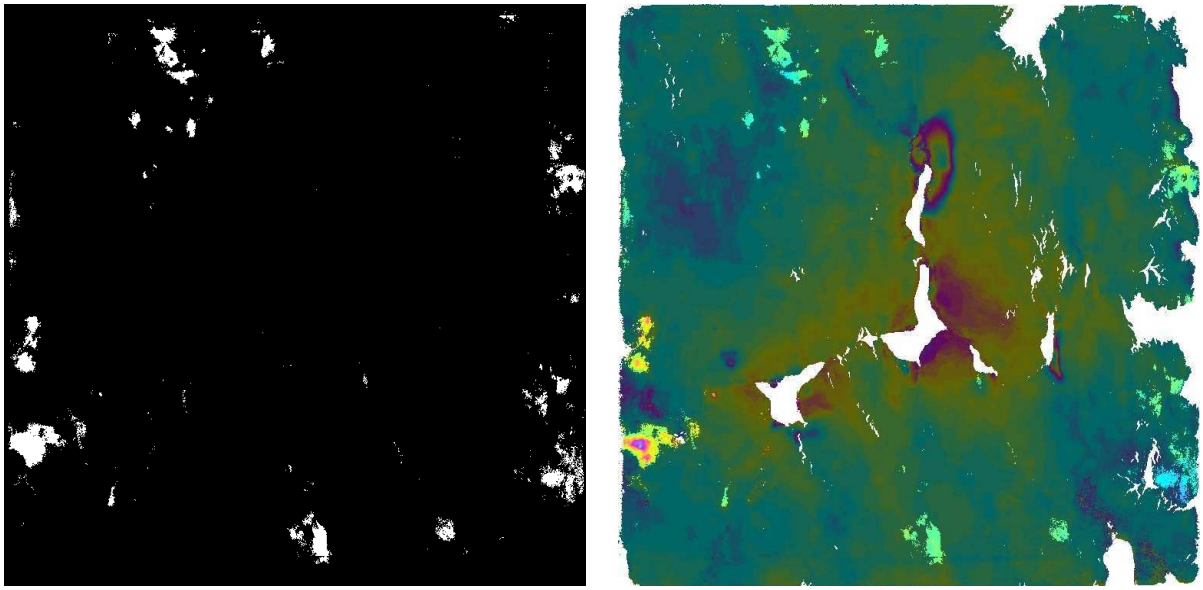
avec la contrainte de CM. Toutes les expériences sur les images de la STIS du lac Mead sont effectuées sur un ordinateur standard (processeur AMD Athlon Dual Core @ 1GHz avec 1 Go de mémoire RAM sous le système d'exploitation Linux noyau 2.6.15-1.2054_FC5 x86_64).

7.1.3 Résultats qualitatifs et interprétations

En réglant s à 3, σ à 10000 ($\sigma_{rel} \approx 2\%$) et κ à 6, 10173 motifs sont obtenus [119]. Pour tenir compte des informations précises, les MSFG avec le maximum d'événements sont sélectionnés. Nous avons trouvé 5 motifs avec 15 événements. Ces motifs sont les suivants :

- motif 1 : 1x15 ($supp = 10636, CM = 6,02$)
- motif 2 : 3x11_1x4 ($supp = 18987, CM = 6,35$)
- motif 3 : 3x10_2_1x4 ($supp = 16655, CM = 6,19$)
- motif 4 : 3x10_1x5 ($supp = 16738, CM = 6,46$)
- motif 5 : 3x9_2_1x5 ($supp = 11035, CM = 6,27$)

Le motif 1 indique que certains pixels sont constamment associés à des valeurs négatives de différence de phase au cours du temps. La localisation de ce motif est présentée dans la Figure 7.6a).

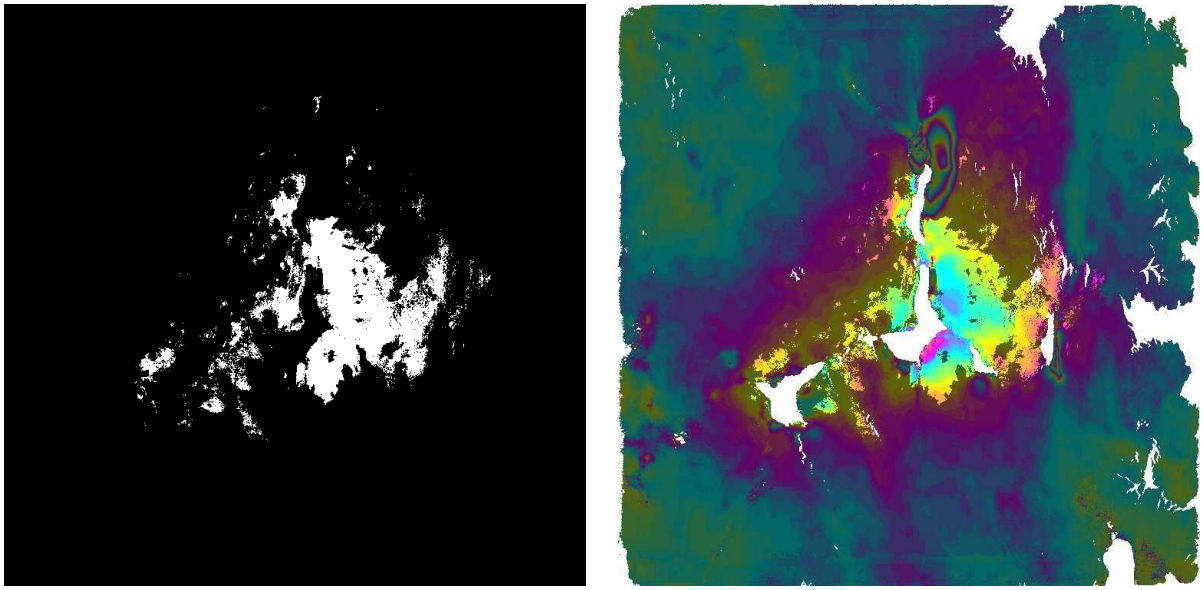


a) b)

FIG. 7.6 – a) Localisation du motif 1 : 1x15 ($supp = 10636, CM = 6,02$) et b) Superposition du motif 1 (zones éclairées) et de la vitesse moyenne de subsidence ou de soulèvement.

Un tel motif peut être dû à une évaluation incomplète des délais de phase atmosphériques de la date maîtresse qui correspond à l'ensemble des 15 dates esclave, ou à des déformations de soulèvement qui affectent tous les 15 dates esclave après 08/10/1995. Pour vérifier la validité de cette dernière hypothèse, la vitesse moyenne de soulèvement ou de subsidence provenant de l'ensemble des données interférométriques (50 images) est calculée. La Figure 7.6b) montre la superposition (zones éclairées) de la vitesse moyenne de subsidence ou de soulèvement et le motif 1. La vitesse de soulèvement ou affaissement est représentée avec une échelle de couleurs enveloppée (rouge / jaune / vert / bleu / violet). Un cycle de couleur positif (respectivement négatif) des zones stables (bords de l'image) à des zones de déformation correspond à une subsidence (respectivement, soulèvement) de 2,2 mm / an. La superposition souligne la principale zone de soulèvement près de Las Vegas (en bas à gauche) qui est probablement due au pompage diminué de l'eau dans cette partie de l'aquifère de Las Vegas, comme observé dans [11].

Les premiers symboles de motifs 2, 3, 4 et 5 indiquent de grandes différences de phase positives par rapport à l'image maîtresse et les derniers symboles indiquent des grandes différences de phase négative. La localisation conjointe de ces motifs est représentée dans la Figure 7.7a). Ces motifs sont en corrélation avec les fluctuations du niveau d'eau qui a augmenté entre 08/10/1995 et 1998, et a diminué après 2000. En d'autres termes, ces motifs suggèrent qu'il devrait y avoir des pixels pour lesquels la subsidence (respectivement, soulèvement) est observée lors de l'augmentation du niveau de l'eau (respectivement diminution). Un tel comportement serait confirmé par un coefficient de régression positif entre les délais de phase et les fluctuations du niveau de l'eau. Pour vérifier cette hypothèse, le coefficient de régression est calculé en utilisant l'ensemble des données interférométriques. Des coefficients de régression positifs sont obtenus pour la localisation de motifs 2, 3, 4 et 5 (voir Figure 7.7b). Le coefficient de régression est représenté par une échelle de couleurs enveloppée (rouge / jaune / vert / bleu / violet). Un cycle de couleur positif (respectivement négatif), des zones stables à des zones de déformation, correspond à une subsidence (respectivement, soulèvement) de 0,7 mm lorsque le niveau d'eau augmente de 1 m.



a) b)

FIG. 7.7 – a) Localisation conjointe des motifs 2, 3, 4 et 5 et b) Superposition de la localisation conjointe des motifs 2, 3, 4 et 5 (zones éclairées) et du coefficient de régression entre les délais de phase et les fluctuations du niveau d'eau.

Tous les motifs se rapportent donc à des déformations du sol et non à des perturbations atmosphériques, ce qui confirme que la notion de MSFG peut être utilisée pour trouver des motifs spatio-temporels décrivant des phénomènes non-aléatoires dans des STIS [119].

7.2 La STIS de Chamonix Mont Blanc - Polarimétrie radar

Parallèlement à l'interférométrie qui est mise en oeuvre en utilisant une polarisation fixe, la polarimétrie RSO permet la discrimination des propriétés intrinsèques de la cible en exploitant différentes polarisations [23].

Si l'interférométrie utilise un couple d'images afin d'en déduire la hauteur d'un centre de phase, la polarimétrie exploite la nature vectorielle du champ électromagnétique et l'interaction avec la cible imagée. La polarisation de l'onde est décrite par la position du vecteur de champ électrique dans le plan d'onde en émission et en réception.

Les vecteurs de champ électrique émis par un système radar peuvent être horizontaux (H) ou verticaux (V). La polarisation du signal reçu en retour dépend de la manière dont le signal a été rétrodiffusé (type de réflexion, matériau et forme de la cible, rétrodiffusion avec deux rebonds ou dans un petit volume). Le signal en retour est donc lui aussi polarisé horizontalement ou verticalement, selon les propriétés polarisantes de la cible rencontrée. On peut distinguer plusieurs canaux selon la polarisation émise et transmise, on accole alors les lettres dans cet ordre. Pour un signal émis verticalement et reçu horizontalement on notera VH. Deux configurations copolaires sont possibles, HH et VV, et deux configurations contrapolaires HV et VH, équivalentes dans le cas monostatique (où l'antenne à l'émission est la même que celle à la réception).

Les informations enregistrées par les capteurs radar polarimétriques sont liées aux paramètres qui décrivent l'orientation et de la forme des diffuseurs présents dans la cellule de résolution et ce grâce aux différents mécanismes de diffusion.

La polarisation VV permet de dégager principalement les effets de capillarité de la surface imagée, comme la surface de l'eau (polarisation choisie pour le satellite ERS dédié en premier lieu à l'observation de la mer). Elle caractérise les zones où la couche de glace est inexistante, fine ou très lisse. En revanche, le rapport de la co-polarisation s'équilibre sur les zones de glace fortement déformées [182]. La polarisation HH est bien adaptée à l'étude de l'humidité du sol, en faisant ressortir les objets de différentes hydrométries (distinction glace/eau). Les polarisations croisées HV et VH mesurent une dépolarisation du signal entre son émission et sa réception, traduisant des rebonds multiples au niveau de sa rétrodiffusion. Ce phénomène correspond par exemple aux textures de végétation très feuillue ou encore au niveau des glaciers à la présence de crevasses, séracs ou zones recouvertes de cailloux. En raison de la rétrodiffusion minimale pour l'eau, ces types de polarisations contrapolaires peuvent être utilisés pour améliorer la discrimination entre eau et glace ou terre (les deux derniers possédant une rétrodiffusion volumique) [4].

Les principaux apports de la polarimétrie sont la description et la caractérisation des cibles, et leur classification. Grâce à la polarimétrie, l'interférométrie RSO devient plus sensible à la distribution des objets orientés (la végétation ou les zones urbaines par exemple).

7.2.1 Données et pré-traitements des données

Dans cette section, une application des données RSO polarimétriques est proposée [126, 116] : une STIS Radar à Synthèse d'Ouverture Polarimétrique (en anglais Polarimetric Synthetic Aperture Radar) (PolSAR) du satellite RADARSAT-2 couvrant une zone de haute montagne : le site test Chamonix Mont-Blanc du projet EFIDIR. Le jeu préliminaire de données expérimentales [126] est composé de 4 images completèment polarimétriques acquises durant l'hiver 2009, le 29 janvier, le 22 février, le 18 mars et le 11 avril. Essentiellement, les systèmes RSO mesurent à la fois l'amplitude et la phase du signal rétrodiffusé, produisant une image complexe pour chaque enregistrement. Le PolSAR est une extension du système d'imagerie RSO, les capteurs étant en mesure d'émettre et de recevoir deux polarisations, généralement horizontale (H) et verticale (V).

RADARSAT-2 est le satellite RSO commercial canadien de prochaine génération qui a été lancé en décembre 2007 sur un véhicule Soyouz depuis le cosmodrome russe de Baïkonour, au Kazakhstan [5]. Il a été placé sur la même orbite que RADARSAT-1, qu'il suit à 50 minutes d'intervale (orbite inclinée à 98,6° d'altitude de 798 km). Il offre de puissantes capacités techniques novatrices qui permettent de faciliter la surveillance maritime, la surveillance des glaces, la gestion des catastrophes, la surveillance environnementale, la gestion des ressources ainsi que les activités de cartographie. Le capteur RSO fonctionne dans la bande C ayant une longueur d'onde de 5,55 cm et une polarisation quadruple (HH, HV, VH et VV).

Les images PolSAR peuvent assurer différentes mesures sur la base de la spécificité radar : la pénétration de la neige et de la glace qui conduit à une observation en dessous de la surface et les ondes cohérentes polarisées qui permettent aux médiums de rétrodiffusion d'être analysés. Toutefois, les acquisitions polarimétriques RSO fournissent des images complexes multi-canaux, qui nécessitent une chaîne de traitement compliquée, en particulier dans le contexte des glaciers alpins de haut relief, avec des mouvements et des changements de surface rapides et une signature RSO plutôt méconnue du mélange glace / neige / roches. Une étape de pré-traitement est nécessaire afin d'obtenir des informations de niveau plus haut que les données RSO initiales (des détails supplémentaires sont fournis dans [126]).

Les premières acquisitions RSO polarimétriques résultent dans une image complexe avec 4 canaux, chaque canal correspondant à une configuration de polarisation différente, habituellement désigné par S_{HH}, S_{VV}, S_{HV} et S_{VH}. La première étape de traitement consiste à transfor-

mer les canaux complexes en 3 canaux (en configuration monostatique $S_{HV} \backsimeq S_{VH}$) exprimés dans la base de Pauli [47], qui est généralement préférée pour obtenir un vecteur cohérent de rétrodiffusion $[k]$ plus proche des phénomènes physiques de rétrodiffusion des ondes :

$$[k] = \frac{1}{\sqrt{2}} \begin{bmatrix} S_{HH} + S_{VV} \\ S_{HH} - S_{VV} \\ 2S_{HV} \end{bmatrix}. \tag{7.1}$$

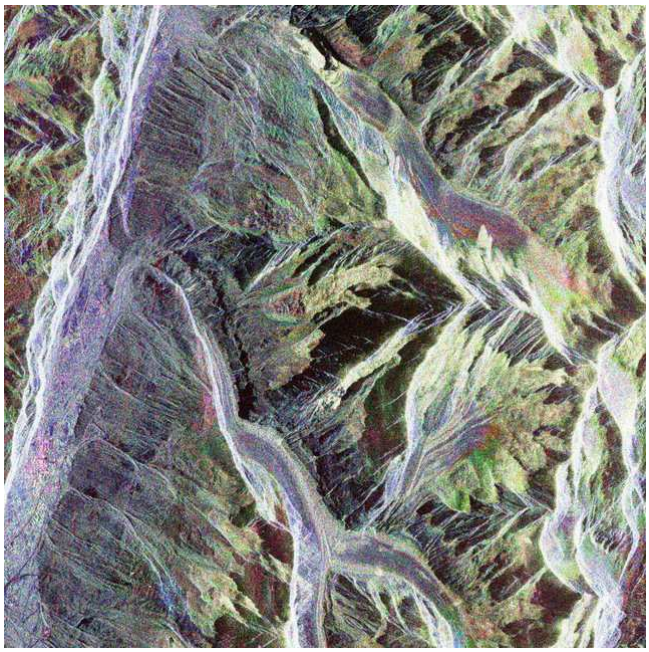


FIG. 7.8 – Image RADARSAT-2, 29/01/2009, région du Mont-Blanc ; la composition en couleurs des 3 amplitudes dans la base Pauli, R : HH-VV, V : 2HV, B : HH + VV, 2048 × 2048 pixels.

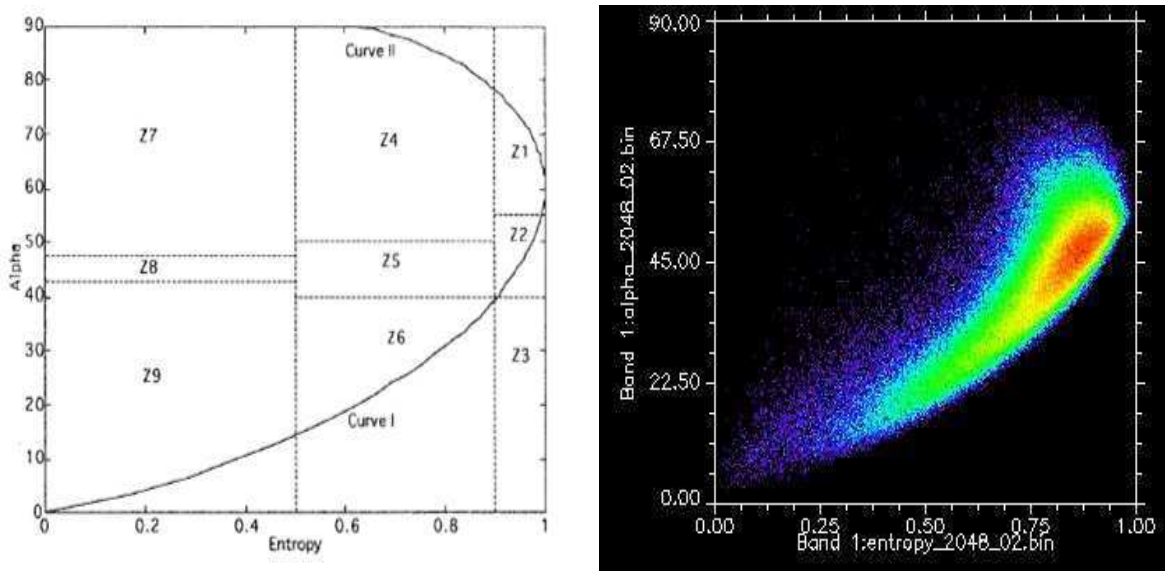
La Figure 7.8 illustre les 3 amplitudes dans la base de Pauli d'une image RADARSAT-2 acquise sur la zone Chamonix Mont-Blanc. Cette sous-image comprend 3 glaciers (Argentière, Talèfre et Mer-de-Glace), la vallée de Chamonix à environ 1000 m d'altitude (à gauche vers le bas) et des montagnes jusqu'à 4000 m.

Le phénomène de chatoiement (speckle) affecte les cibles réparties et rend difficile le travail directement avec les vecteurs de rétrodiffusion à l'exception du cas d'une cible cohérente. La deuxième étape de pré-traitement consiste à estimer la matrice de cohérence hermitienne 3×3. Dans cette expérience, le filtre Lee classique est appliqué pour réduire l'effet du bruit de chatoiement (speckle) sur la matrice de cohérence.

La troisième étape de pré-traitement consiste à appliquer la décomposition de Cloude et Pottier [47] pour exprimer la matrice de cohérence comme une somme pondérée de trois matrices représentant 3 mécanismes de rétrodiffusion pure. Plusieurs caractéristiques PolSAR sont généralement dérivées de cette décomposition pour discriminer les différents mécanismes de rétrodiffusion.

Parmi les caractéristiques PolSAR, l'entropie H et le paramètre α indiquent le comportement aléatoire de la rétrodiffusion globale et le mécanisme moyen de rétrodiffusion de la surface à rétrodiffusion double rebond. Elles sont fortement liées aux propriétés géophysiques de la zone cible au sol fournissant des informations fiables pour une autre classification. Dans [48], neuf zones de regroupement sont proposées pour décrire le plan H et α. Les frontières de ces

clusters nets prédéfinis illustrées sur la Figure 7.9a) sont souvent utilisées comme références pour interpréter les espaces d'attributs en termes de mécanismes de rétrodiffusion ou pour initialiser des techniques de regroupement (clustering).



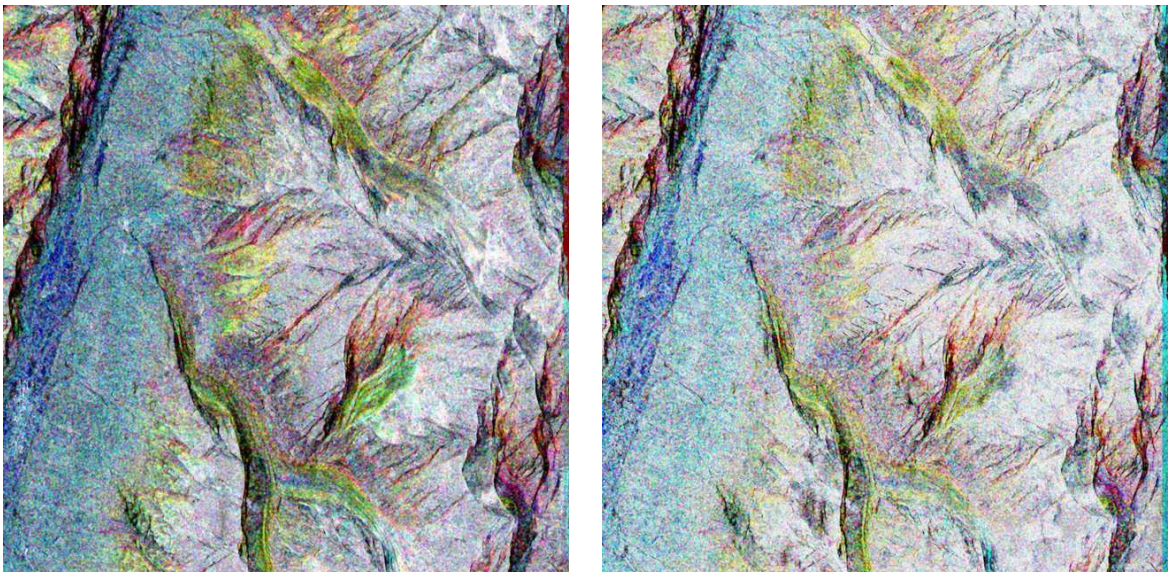
a) b)

FIG. 7.9 – L'espace de la caractéristique H - α ; a) la partition en 9 zones correspondant aux différents types de rétrodiffusion [48] ; b) Distribution de l'image RADARSAT-2 22/02/2009 sur la zone du Mont-Blanc.

Pour suivre une chaîne classique de pré-traitement PolSAR, les caractéristiques (H, α) sont utilisées comme des informations d'entrée. La distribution 2D de ces caractéristiques sur la zone de Mont-Blanc est illustrée dans la Figure 7.9b). L'information «temporelle» véhiculée par ces caractéristiques peut être observée en combinant 3 dates différentes dans une composition de couleurs RVB (Figure 7.10). Les résultats montrent que des changements significatifs ont eu lieu entre les 3 dates, et que les évolutions sont liées à des différentes parties des images : les glaciers, les montagnes et la Vallée de Chamonix apparaissent avec des couleurs différentes, qui varient également avec l'orientation ou l'hauteur de la pente. Cette analyse visuelle est essentiellement qualitative et limitée à 3 dates affectées aux couleurs R, V et B.

L'extraction des MSFG exige que les séquences d'entrée de valeurs de pixels soient transformés en séquences de symboles. La stratégie de quantification doit conserver les informations utiles. Dans le cas des données PolSAR, les deux canaux d'entrée pour chaque date sont la valeur d'entropie dans [0, 1] et la valeur d'angle α dans [0, 90°]. Les deux valeurs peuvent être quantifiées dans un nombre réduit d'intervalles, résultant dans un partitionnement régulier de l'espace H - α. Au lieu de cette quantification générale, l'approche proposée consiste à utiliser la partitionnement de l'espace H - α illustré dans la Figure 7.9a). Cette représentation symbolique fournie par le domaine PolSAR rend l'interprétation des MSFG extraits plus facile pour les utilisateurs finaux.

Les expériences ont été effectuées sur 4 images RADARSAT-2 disponibles sur le site test Chamonix Mont-Blanc en appliquant d'abord la chaîne de pré-traitement présentée auparavant. Ensuite, les images de H et α résultantes (4 paires d'images de 2048 × 2048 pixels) sont explorées en codant les positions de pixels en fonction des 9 zones.



a) b)

FIG. 7.10 – Composition en couleurs de a) l'angle α et b) l'entropie ; R : 2009/01/29, V : 2009/03/18, B : 2009/04/11.

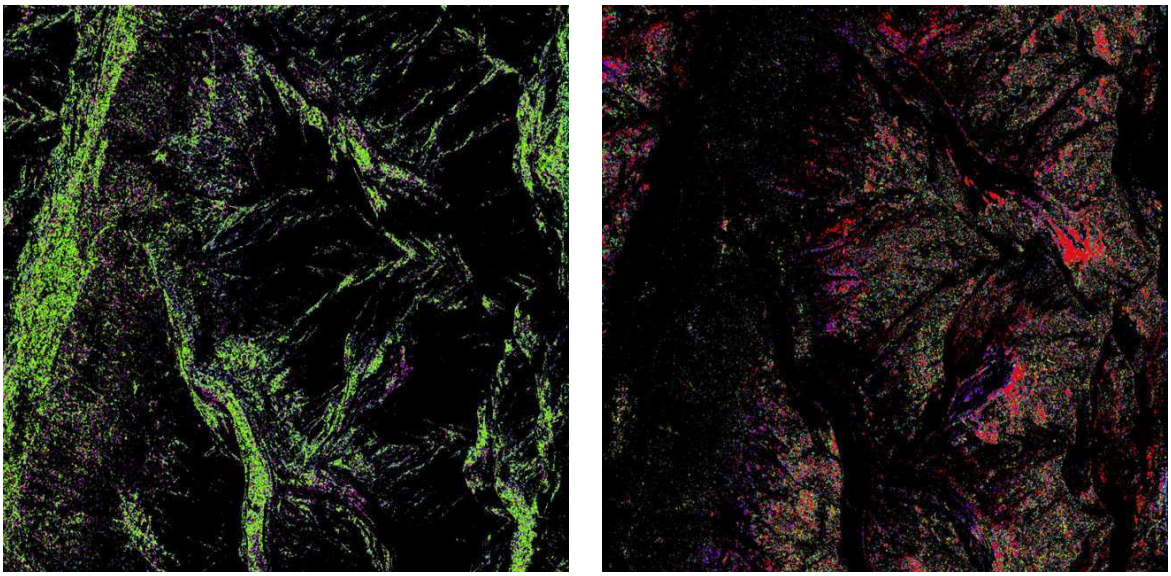
7.2.2 Résultats préliminaires

En recherchant les MSFG par l'application de la contrainte de CM à l'aide de la relaxation avec la contrainte de CRSM ($\kappa = \mu$) les paramètres d'extraction ont été établis en accord avec les désirs des experts (utilisateurs). Avec un seuil de surface minimale σ de 10000 et un seuil minimum de connexité moyenne κ fixé à 4 sont trouvés 14 MSFG différents. Parmi eux, 6 motifs SFG retiennent l'attention :

- $6 \rightarrow 6 \rightarrow 6$, qui apparaît souvent sur les parties inférieures du glacier et sur la moraine environnante,
- $4 \rightarrow 4$, qui semble être complémentaire au motif précédent en apparaissant sur les parties supérieures du glacier,
- $5 \rightarrow 5$, qui apparaît sans localisation dominante. Ceci peut être expliqué par la position centrale de cette zone dans les distributions H - α observée sur les images entières (cf. Figure 7.9),
- $5 \rightarrow 6$ et $6 \rightarrow 5$, qui apparaissent également sans localisation dominante. Ceci peut être interprété par la frontière entre les zones 5 et 6 qui sépare les points qui ont un mécanisme de rétrodiffusion proche de cette frontière. La variance d'estimation H, α détermine nombreux pixels de changer la valeur entre 5 et 6 ou à l'inverse au moins une fois dans leur évolution temporelle,
- $6 \rightarrow 9$, qui apparaît souvent sur les versants de la montagne affectés par du recouvrement.

Les deux premiers MSFG sont illustrés dans la Figure 7.11. Les pixels où le MSFG se produit apparaissent avec une couleur qui dépend de ses dates d'occurrence. Cette visualisation permet à l'utilisateur final d'observer pour chaque motif où et quand il apparaît (voir l'annexe B.2). Par exemple, le motif $6 \rightarrow 6 \rightarrow 6$ affiché dans la Figure 7.11a) apparaît la plupart du temps avec la même couleur verte qui correspond à la zone 6 présente dans les images 1, 2 et 3. C'est conforme à l'analyse visuelle des 4 dates sur les glaciers : la date 4 est en avril et les températures de printemps commencent à transformer la couverture de neige du glacier au plus basses altitudes.

En ce qui concerne le coût de calcul, sur un ordinateur portable PC standard (processeur Intel Core 2 Duo @ 2GHz avec 2 Go de mémoire RAM sous le système d'exploitation Linux



a) b)

FIG. 7.11 – Localisation spatio-temporelle des MSFG détectés dans la série temporelle H-α de 4 dates. (A) : $6 \rightarrow 6 \rightarrow 6$; (b) : $4 \rightarrow 4$; différentes couleurs correspondent à différentes localisations temporelles du MSFG.

noyau 2.6.31), ces résultats sont obtenus en moins de 32 s en utilisant moins de 1,6 Go de RAM.

Malgré un petit nombre d'images, ces premiers résultats montrent le potentiel de l'approche de fouille de données d'extraire de MSFG à partir des séries temporelles d'images PolSAR [126, 116]. Des motifs spatialement cohérents et non contraints temporellement peuvent être extraits automatiquement. Ils fournissent à l'utilisateur final une description claire des principales évolutions des mécanismes dominants de rétrodiffusion, associée à leurs localisations spatio-temporelles.

Les orientations des travaux futurs comprennent l'utilisation des MSFG pour analyser des séries temporelles PolSAR plus longues et l'amélioration de l'étape de pré-traitement en utilisant des filtres adaptatifs pour réduire la variance d'estimation H - α en préservant les caractéristiques spatiales.

Conclusion

L'application de cette approche d'extraction à des différents types de données dérivant de différentes STIS et techniques satellitaires vérifie qu'elle bénéficie de la généricité du concept de MSFG. Les STIS utilisées concernent une zone agricole de plaine (ADAM), un lac d'accumulation (lac Mead) et une zone montagneuse avec glaciers (Chamonix Mont Blanc). La nature de données est également diverse : optique en 3 bandes pour le premier site et radar (amplitude, phase) pour des observations interférométriques et polarimétriques, pour les cas suivants. Les expériences sur plusieurs ensembles de données synthétiques et réels ont été réalisées pour vérifier et évaluer l'approche proposée. Les résultats de ces expériences ont suivi les attentes des spécialistes et ont aidé au réglage des paramètres d'algorithmes d'extraction.

Le travail présenté dans ce mémoire a porté sur la découverte de motifs locaux contraints dans des bases de données séquentielles d'images satellitaires. On a privilégié la généricité des contraintes traitées en adoptant une démarche reposant sur des conditions suffisantes pour pousser les contraintes profondément dans l'extraction. Les contraintes utilisées, de support (surface minimum) et de connexité, offrent une forte expressivité pour décrire le type de connaissance à cerner.

Les résultats obtenus sur les motifs contraints permettent de traiter différents problèmes applicatifs réels et valident des connaissances du domaine tout en les quantifiant.

Les résultats obtenus montrent l'importance de la sélectivité des MSFG pour la maîtrise des temps d'exécution et la réduction du nombre de solutions fournies à l'utilisateur. Les différentes expérimentations menées dans la partie III confirment ces observations pour les deux contraintes anti-monotones, de support et de connexité relative au support minimum. De même, l'impact de la relaxation anti-monotone, la plus efficace et efficiente méthode proposée, se ressent particulièrement lorsque la sélectivité de la contrainte est forte (les seuils de contraintes sont élevés).

Dans la littérature, on insiste sur le fait qu'un processus ECD de qualité requiert une interactivité et une itérativité fortes avec l'utilisateur/analyste. L'interactivité du processus doit mettre en avant l'utilisateur au sein de l'extraction.

L'itérativité se traduit par la répétition du processus. De cette manière, l'utilisateur peut ajuster les paramètres du processus de fouille. Par exemple, il doit pouvoir modifier ou compléter la contrainte. Le chapitre 6 est un bon exemple. Il montre comment les modifications des différents paramètres d'extraction dont le choix est fait par l'utilisateur (le nombre de symboles pour la discrétisation, les niveaux des seuils pour les contraintes, etc.) peuvent affecter l'extraction dans le sens désiré par celui-ci. Parmi les paramètres de sortie il y a également ceux d'intérêt pour l'utilisateur : le nombre de motifs extraits et le temps d'exécution.

La démarche heuristique explore, en augmentant la complexité, les étapes d'extraction des motifs séquentiels sans aucune contrainte, des MSF avec la contrainte de support et de MSFG avec différents types de contraintes de connexité. Ainsi, l'étude quantitative a mis en évidence l'influence des types de contrainte(s) appliquée(s) sur les paramètres de sortie. Les

caractéristiques du passage d'un type d'extraction à l'autre sont mises en évidence, également par une équivalence avec l'action d'un filtre dont la fonction de transmission est figurée. Les motifs extraits sont analysés de façon quantitative et qualitative. Apparemment, l'intérêt de l'utilisateur est d'avoir à sa disposition, après une extraction, un nombre réduit de motifs longs en correspondance avec son pouvoir d'interprétation et traitement. En fait, ce nombre doit être suffisamment grand pour décrire la palette d'entités qui couvrent la scène étudiée et la diversité de situations dans lesquelles ces entités peuvent se trouver. D'autre part, l'obstination pour obtenir et interpréter uniquement des motifs très longs, soient-ils maximaux ou fermés, est tempérée par les résultats de l'étude qualitative de chapitres 6 et 7. Là, on voit que le rapport entre la couverture d'un motif et la couverture d'une entité potentiellement correspondante souffre d'une décroissance rapide avec la spécialisation du motif. Contrairement aux apparences, il n'y a pas une règle générale assurant que la spécialisation assure une amplification de la pureté (conformité d'une description). Dans la désignation des bons candidats pour un éventuel regroupement ou classification, seul l'expert peut décider sur le compromis couverture - pureté, toutes les deux affectées par la longueur du motif. Ainsi, on trouve que la longueur grande d'un motif n'est pas le critère essentiel pour cette désignation. D'ailleurs, il y a de nombreux exemples de motifs courts et de longueur intermédiaire qui ont une couverture excellente et un degré élevé de justesse de la description d'entités plus générales. La considération des motifs incomplets comme longueur permet d'ignorer une ou quelques valeurs de la série temporelle. De cette façon, les valeurs déviantes (nuage, bruit) sont automatiquement ignorées. De plus, seuls les pixels déviants sont négligés, ce qui permet d'utiliser toutes les données disponibles, sans supprimer des images bruitées.

Un autre aspect intéressant est le choix que l'analyste ou l'utilisateur doit faire sur le nombre de symboles utilisés pour la discrétisation des valeurs de pixels. Les exemples présentés montrent que, dans le cas de la réduction du nombre de symboles, l'augmentation évidente de la couverture compense la possible diminution de la précision de description. De même, a été mis en évidence un comportement intéressant pour les longueurs et les degrés de connexité élevés des motifs : le nombre de motifs augmente lorsque le nombre de symboles diminue.

L'analyse atteste la généricité du concept de MSFG et la pertinence de la démarche d'extraction proposée. Les images de localisation spatiale et temporelle des motifs et de leurs variantes temporelles sont comparées avec une vérité terrain ou d'autres types de connaissances spécifiques du domaine. Les motifs extraits confirment leur capacité à décrire avec une bonne précision les entités de la couverture des sols, objets et phénomènes, leur localisation spatiale et temporelle et aussi leur évolution. Dans le cas de la STIS ADAM (chapitre 6), l'étude trouve la qualité de MSFG extraits et la possibilité d'assurer un bon compromis entre les couvertures thématiques et les puretés de description. Dans le chapitre 7, pour les données interférométriques, les MSFG extraits décrivent correctement les déformations de la croûte terrestre suivant les variations du niveau de l'eau du lac et confirment leur capacité de trouver des modifications produites par des phénomènes non-aléatoires. Dans le cas de données polarimétriques, l'étude met en évidence le potentiel du concept de MSFG et de l'approche d'extraction pour fournir une description claire de principales évolutions des mécanismes de rétrodiffusion et leurs localisations spatio-temporelles.

Chapitre 8

Bilan et perspectives

L'extraction automatique de connaissances à partir d'images satellitaires dans un contexte spatio-temporel est un défi majeur dans le domaine de la télédétection. Dans ce contexte, nous proposons une nouvelle technique pour l'extraction des motifs d'évolution d'une base de séquences correspondant aux données de séries temporelle d'images satellitaires. Notre objectif est de vérifier et valider l'applicabilité d'une technique d'extraction basée sur un concept nouveau, le motif séquentiel fréquent groupé.

Le mémoire est la continuation des travaux concernant l'introduction des techniques de fouille de données dans l'étude des STIS, par l'extraction de motifs séquentiels fréquents, sur une ou plusieurs bandes, réalisée dans des travaux antérieurs ([124, 123], annexe C).

L'analyse des données images qui est faite au niveau pixel préserve le traitement de l'information au niveau de la résolution d'observation et est indépendante de l'application. La première caractéristique principale est l'utilisation de l'évolution temporelle des valeurs des pixels pour décrire, caractériser et discriminer les comportements des pixels. La détermination d'évolutions temporelle est faite au long de toute la période d'observation sans aucune discrimination a priori et les informations obtenues sont plus complexes en contenu en comparaison avec une simple détection de changements. De plus, l'étude est réalisée sur des images entières de grandes dimensions afin de contenir l'intégralité des informations spatiales. La démarche assure non seulement la mise en évidence mais également la caractérisation des changements observés, en fournissant les motifs partagés par ces ensembles de séquences. De cette manière, devient possible la réalisation d'une carte finale de localisation des types d'évolutions pertinentes pour l'utilisateur. Cette image synthétique pourrait aider dans les processus de compression et d'indexation des STIS. La démarche est non-supervisée ; elle permet l'analyse de la totalité des données fournies par la série d'images, sans une sélection d'objet ou de classe a priori.

La deuxième caractéristique principale est la considération des caractéristiques spatiales des données et leur introduction dans le processus d'extraction de motifs séquentiels de STIS, comme expressions de la connaissance du domaine, en concordance avec les attentes de l'utilisateur. Ainsi, dans la Partie II, en plus de la contrainte de surface minimum, sont introduites des mesures de connexité des pixels couverts par le même motif (la connexité globale, CG, moyenne, CM, et relative au support minimum, CRSM). Sur la base de ces mesures sont établies des contraintes de connexité minimum permettant une implantation active dans le processus d'extraction. C'est la troisième caractéristique principale de la démarche de la thèse et elle est concrétisée par l'utilisation des différentes contraintes de connexité basées sur les mesures de connexité introduites, comme des conséquences de la propriété de dépendance spatiale des caractéristiques des motifs d'évolution. La contrainte sur la CM a une signification assez claire pour

l'utilisateur et sur sa base est défini un nouveau concept, le motif séquentiel fréquent groupé, MSFG, qui répond à l'intérêt de l'utilisateur et dont la généricité est attestée dans les différentes expérimentations. Les mesures de CG et de CRSM sont anti-monotones et permettent donc la réalisation de contraintes qui supportent d'être «poussées» activement au sein du processus d'extraction avec de bons résultats sur le plan de l'efficience et de l'efficacité. Différentes approches par rapport à l'utilisation passive ou active des contraintes, seules ou en conjonction, dans le processus d'extraction de motifs sont étudiées. Parmi celles-ci, la relaxation optimale de la contrainte sur CM par la contrainte sur CRSM a les meilleurs résultats pour une extraction complète et efficiente de MSFG [119, 121, 122].

En effet, notre approche propose un équilibre dans l'utilisation des trois dimensions du pixel de STIS : radiométrique, temporelle et spatiale. La considération de la caractéristique d'évolution comme élément descriptif et discriminant augmente l'importance de la dimension temporelle. La dimension spatiale bénéficie de la considération de la caractéristique spatiale de connexité comme outil essentiel pour l'extraction des MSFG.

Dans cette approche, du fait de l'hétérogénéité des motifs spatio-temporels observés, l'extraction est orientée vers des motifs locaux, comme sources d'information pour des développements ultérieurs, et pour assurer une description spatialement et temporellement localisée plutôt qu'une description globale de la scène. Les motifs locaux contribuent à la réalisation des modèles descriptifs (par clustering, classification, etc.).

Les exigences d'un processus ECD de qualité, à savoir une interactivité et une itérativité fortes avec l'utilisateur/analyste sont assurées par la possibilité d'ajuster plusieurs paramètres d'entrée, de choisir le type de contrainte utilisée ou même d'accéder aux différents types d'extraction. L'approche proposée est vérifiée et validée par des applications sur des données réelles, exposées dans la Partie III. Les expérimentations attestent de la pertinence de l'approche pour l'extraction de motifs d'intérêt, la signification des motifs d'évolutions extraits et la généricité du concept de MSFG.

Dans la Partie III, les étapes d'extraction de motifs sans contrainte MS, des MSF avec la contrainte de support seule, et des MSFG avec l'ajout des différents types de contraintes sur connexité sont enchaînées. Le passage d'un type d'extraction à l'autre est assimilé à l'action d'un filtre équivalent dont la fonction de transmission est présentée. Les expériences montrent comment ajuster les paramètres d'entrée (le nombre de symboles pour la discrétisation, les niveaux des seuils pour les contraintes) et comment se servir de différents types de contraintes employés et des types d'extraction développés pour obtenir les caractéristiques des motifs extraits attendues par l'utilisateur.

L'étude quantitative est complétée par une étude qualitative des résultats de l'extraction. Les principaux paramètres étudiés sont la couverture en pixels et le degré de conformité de la description (la pureté) assurés par les motifs extraits. Dans la désignation des bons candidats pour un éventuel regroupement ou classification, seul l'expert peut décider sur le compromis couverture - pureté, les deux étant affectées par la longueur du motif. Ainsi, on trouve que la longueur d'un motif n'est pas le seul critère pour cette désignation. D'ailleurs, il y a de nombreux exemples de motifs courts et de longueur intermédiaire qui ont une couverture excellente et un degré élevé de justesse de la description d'entités plus générales. La prise en compte de motifs incomplets (dont la taille est inférieure au nombre d'images de la série) permet d'ignorer une ou quelques valeurs de la série temporelle. De cette façon, les valeurs aberrantes (nuage, bruit, étalonnage du capteur) sont automatiquement ignorées. De plus, seuls les pixels déviants sont négligés, ce qui permet d'utiliser toutes les données disponibles, sans supprimer les images bruitées. Ces faits attestent d'une généricité de l'approche et correspondent à une certaine robustesse au bruit.

Dans le cas de motifs très longs, un autre aspect intéressant est offert par la réduction du nombre de symboles utilisés pour la discrétisation des valeurs des pixels, pouvant aller jusqu'à la binarisation de l'image ; elle peut avoir des effets bénéfiques comme la croissance de la couverture, de la connexité et même du nombre de motifs.

Même dans le cas d'un échantillonnage temporel irrégulier, l'approche d'extraction développée permet la détection de zones compactes et les motifs extraits confirment leur capacité à décrire avec une bonne précision les entités de la couverture des sols, objets et phénomènes, leur localisation spatiale et temporelle et aussi leur évolution. Pour les motifs de longueur incomplète, la technique de localisation temporelle permet de déceler des entités qui apparaissent ensemble dans l'interprétation d'un motif.

Les MSFG peuvent être une bonne solution pour la caractérisation préliminaire des données séquentielles et ils donnent des ensembles de motifs satisfaisants en compacité et en représentativité pour une application particulière, permettant une exploration directe et efficace des tendances d'évolution [120, 126, 119, 160, 116, 121, 122].

Les résultats d'extraction de MSFG ne sont pas finaux ou suffisants mais ils permettent la réalisation d'une description préliminaire de STIS de point de vue des évolutions temporelles de la radiométrie des zones correspondantes aux pixels d'images, et du point de vue de la surface et de la connexité des pixels couverts par ces évolutions. Les MSFG découverts via le processus d'extraction sont intéressants non seulement en eux-mêmes, mais ils peuvent également être utiles pour l'analyse d'autres données et tâches d'exploration.

Perspectives

Les prochaines directions de recherche concernent : a) le raffinement de la démarche actuelle, b) l'implantation de la méthodologie dans des systèmes de caractérisation complète des données séquentielles et c) la diversification de la palette d'applications.

Concernant l'amélioration des éléments de la démarche, il y a l'intention d'introduire la possibilité de choisir la dimension du voisinage pour le calcul de la connexité locale et une augmentation du rôle de celle-ci. Les mesures de connexité utilisées actuellement, connexité moyenne et relative au support minimum, ont un caractère global en considérant toutes les occurrences d'un motif. On propose l'utilisation d'une mesure de type «relative au support» pour chaque occurrence (pixel) du motif examiné de sorte que, en correspondance avec un seuil, les pixels faiblement liés ou isolés puissent être éliminés. Cela permettrait une prise en compte locale de l'aspect spatial. Concernant les caractéristiques temporelles, on envisage la considération de la distribution réelle des dates d'acquisition et non seulement leur ordre comme on fait en présent.

Un autre aspect de l'amélioration est l'application, dans le processus d'extraction de la seule contrainte anti-monotone sur connexité générale, CG, (sans la contrainte préalable de support CS). Comme est montré dans le chapitre 4, la mesure de CG correspond au produit du support par la connexité moyenne. L'intention est de récupérer des MSFG qui ont une bonne connexité mais ne passent pas la condition de support (la zone du triangle ABH dans la Figure 5.1).

Du point de vue du post-traitement prévu pour raffiner les motifs extraits, l'intention est de développer des techniques (par exemple similaires à celles utilisées pour les motifs de longueur complète présentées dans l'annexe C), dans le but de réaliser un clustering ou une classification après la qualification d'un expert. Le fait que les MSFG de puretés élevées soient de bons candidats pour la caractérisation des entités terrestre va être exploité par l'élaboration des techniques de comparaison des motifs de longueur différents et par le développement des stratégies coopératives de classification.

La description de la STIS par ses évolutions constitue un ensemble d'informations en soi,

qui permet à l'expert de qualifier les motifs extraits. Ces informations peuvent être réinjectées comme des connaissances dans un processus plus global d'étude de la scène, par exemple, le cadre générique LeGo [132], qui utilise des techniques existantes d'extraction de motifs locaux pour une modélisation globale dans une variété de tâches de fouille de données.

Concernant les nouvelles applications, l'intention est d'utiliser cette approche à des données de tomographe 3D, la dimension d'ordre étant l'un des axes spatiaux.

Quatrième partie

Annexes

Annexe A

Pré-traitements des données

Hormis le pré-traitement effectué par le fournisseur (voir la sous-section 6.1.1), les données ont supporté des préparations spécifiques avant l'application de la fouille de données, des préparations qui consistent en un ensemble d'opérations effectuées sur les données afin d'améliorer leur qualité, et, par conséquent, les résultats de la fouille.

La première opération de pré-traitement, un *nettoyage de données*, a été constitué pour la STIS ADAM par le choix de 20 images les moins bruitées parmi le 39 disponibles, ayant comme critère une présence minimum des nuages et l'intégralité des images. Cette opération n'est pas toujours obligatoire : comme évoqué dans la partie III, la méthode proposée assez robuste au bruit.

Parfois, les attributs existants ne sont pas en mesure de bien refléter les caractéristiques du domaine et la construction de nouveaux attributs peut aider à avoir un nouvel aperçu de la nature du problème. Cette construction est généralement obtenue par la combinaison d'attributs existants, et dans ce travail c'est le cas de l'IVDN qui a été introduit pour mieux décrire et mettre en exergue l'évolution de la végétation (voir la sous-sous-section D.4).

A.1 Réduction du nombre de valeurs des pixels

Dans l'acception la plus simple, les images sont des signaux bidimensionnels - des fonctions dépendantes de deux variables, les coordonnées spatiales. Les valeurs d'une image satellitaire proviennent de l'acquisition à partir d'une scène réelle d'une grandeur physique d'intérêt par un capteur spécialisé. Les images satellitaires sont représentées par un nombre fini de bits, qui résulte de l'échantillonnage et de la quantification du signal continu pris par le capteur. Dans le cas le plus simple, on peut considérer une image numérique comme une matrice dont les éléments, les pixels, ont les valeurs de la fonction image. Ainsi, l'image numérique est constituée par des pixels, chacun pixel étant caractérisé par une position et une valeur. Les valeurs des pixels d'une image satellitaire, acquises par un canal spectral, codent la réflectance provenant des points de la scène entre une valeur minimale nulle (l'absence du signal) et une valeur maximale $M-1$ déterminée par le nombre de bits utilisés pour la représentation binaire des valeurs ($M = 2^B$, où B est le nombre de bits).

L'*histogramme* d'une image est la fonction $h(i)$ définie sur l'intervalle de valeurs quantifiées des pixels $0, 1, ..., M-1$ et à valeurs entières non-négatives, associant à chaque valeur quantifiée

i le nombre de pixels ayant cette valeur.

$$h(i) = \frac{N_i}{L \times C} \tag{A.1}$$

où N_i est le nombre de pixels ayant la valeur i et L et C sont les dimensions de l'image (lignes et colonnes). Cette fonction étant assimilable à une densité de probabilité, l'histogramme vérifie la condition de normalisation

$$\sum_{i=0}^{M-1} h(i) = 1 \tag{A.2}$$

L'histogramme offre des informations sur le contenu de l'image : la gamme de variation des valeurs, la qualité perceptuelle, le nombre de types d'objets. Cette distribution étant de premier ordre, l'histogramme ne décrit pas la répartition spatiale ; il est possible que des images avec un contenu différent soient décrites avec des histogrammes semblables. L'histogramme cumulatif d'une image est la fonction $H(i)$ définie sur l'intervalle de valeurs quantifiées des pixels $0, 1..., M-1$ et à valeurs entières non-négatives, associant à chaque valeur i le nombre de pixels de l'image ayant une valeur inférieure ou égale à i. On a donc :

$$H(i) = \sum_{j=0}^{i} h(j) \tag{A.3}$$

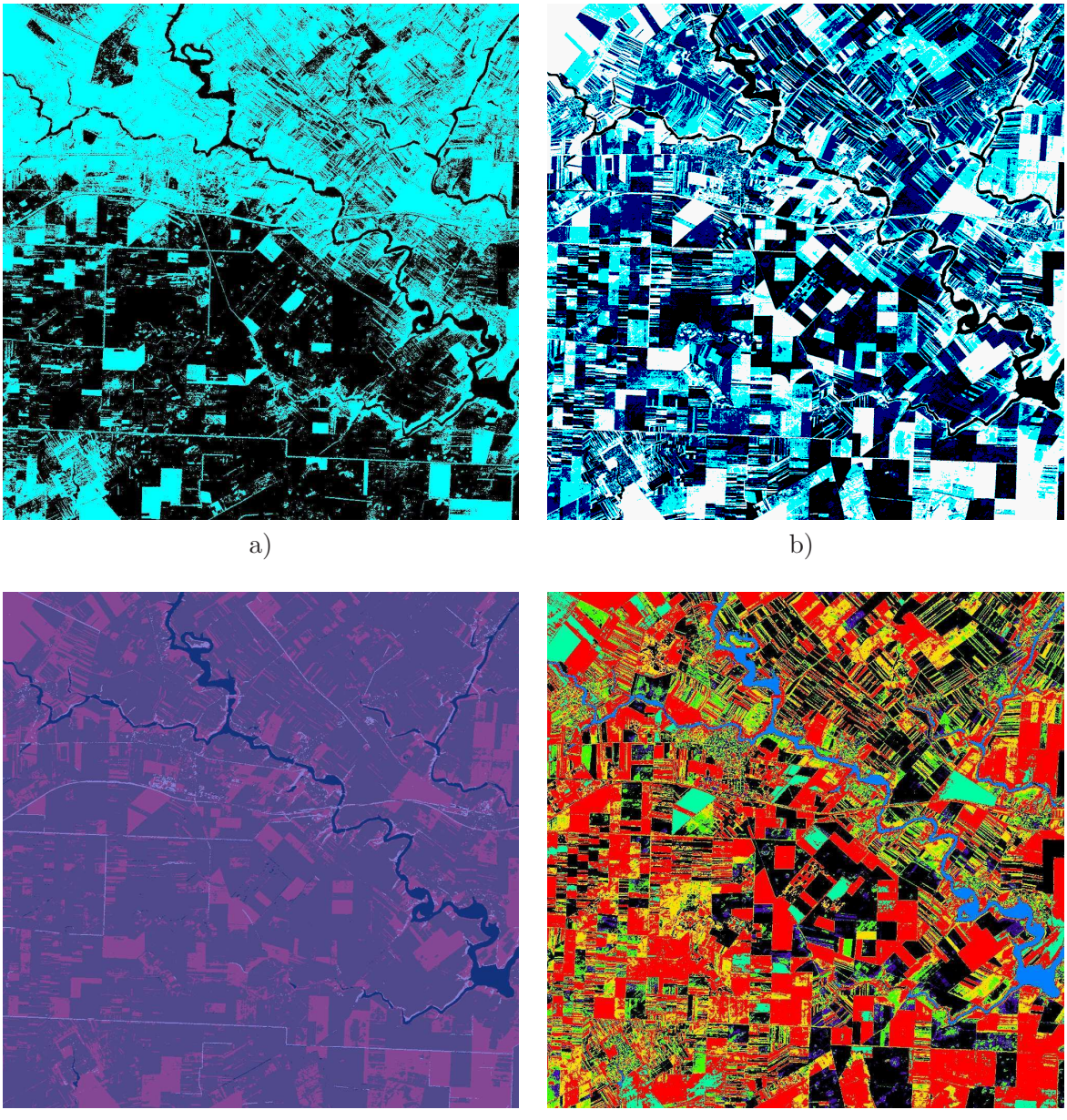
où $i = 0, 1, 2, ..., M-1$.

L'histogramme cumulatif a été utilisé pour réduire le nombre de valeurs des pixels, en le comprimant par une nouvelle quantification. La quantification des valeurs implique le remplacement des valeurs des pixels appartenant à un intervalle par une valeur d'un ensemble fini de valeurs quantifiées. Si la gamme de valeurs initiales des pixels est limitée dans l'intervalle $[\xi_{min}, \xi_{max}]$, cet intervalle est divisé à l'aide des seuils de quantification x_i, où $i \in [1, L+1]$, L étant le nombre d'éléments de l'ensemble des valeurs quantifiées. Pour de valeurs x qui satisfont la relation $x_i < x \leq x_{i+1}$, on approxime la valeur x par la valeur quantifiée y_i ($x_1 = \xi_{min}$ et $x_{L+1} = \xi_{max}$).

Nous avons quantifié les valeurs des pixels dans des intervalles non superposés et contigus qui sont également peuplés. Nous avons utilisé cette technique de discrétisation pour les données optiques de la STIS ADAM [117, 125, 120, 119, 121, 122] et pour les données interférométriques radar de la STIS du lac Mead [120, 119] et de la faille Haiyuan, Chine [160, 122]. Afin de minimiser l'influence des erreurs possibles de calibration, la quantification a été faite pour chaque image, en conservant le même nombre d'intervalles. Pour une date donnée d'acquisition, un pixel a été décrit par une seule étiquette qui indique à quel intervalle appartient cette valeur de pixel.

Les résultats d'une quantification avec $s = 2$ et $s = 4$ intervalles/étiquettes sont présentés dans la Figure A.1 a) et b). Le choix du nombre d'intervalles (ou le nombre d'étiquettes) est difficile sans faire d'expérience. Le compromis entre la description détaillée des évolutions et l'amélioration du processus de fouille de données est difficile à établi a priori. Dans les chapitres 6 et 7 nous étudions l'influence du nombre de symboles (étiquettes), s, sur les résultats de l'extraction de motifs.

Un autre type de quantification est représenté par l'utilisation des seuils de quantification donnés par des experts dans le domaine d'application. Dans [124, 123, 126, 116], nous avons utilisé ce type de discrétisation des données. Dans [124, 123], les valeurs des pixels des images provenant de satellites ERS RSO et METEOSAT sont partagées par l'expert selon les connaissances



a) b)

c) d)

FIG. A.1 – Types de quantifications utilisées : a) 2 intervalles avec histogramme cumulatif (binarisation de l'image) dans PIR ; b) 4 intervalles avec histogramme cumulatif dans PIR ; c) 4 classes avec l'algorithme K-moyennes appliqué sur les 3 bandes spectrales ; d) 8 classes avec l'algorithme Espérance-Maximisation appliqué sur les 3 bandes spectrales

du domaine. Dans [126, 116], la quantification des données polarimétriques radar de la STIS Chamonix, Mont Blanc, est réalisée par l'expert après un pré-traitement qui met en évidence des caractéristiques PolSAR, comme l'entropie et un paramètre α indiquant les mécanismes de rétrodiffusion.

D'autres types de quantifications employées ont été obtenues à partir des clusters fournis par des algorithmes de regroupement. Par exemple, nous avons utilisé les résultats des algorithmes K-moyennes et EM dans [146]. Les images des résultats de ces quantifications sont présentées dans la Figure A.1 c) et d).

A.2 Description d'une STIS à l'aide d'une transformée en cosinus discrète DCT

Une approche courante de réduction de la dimensionnalité est l'utilisation de la Transformée de Fourier Discrète (en anglais Discrete Fourier Transform) (DFT) pour transformer une séquence à partir du domaine temps à un point dans le domaine fréquentiel. Choisir les k premières fréquences, puis représenter chaque séquence comme un point dans l'espace de k dimensions, permet d'atteindre cet objectif. La DFT a une propriété attrayante : l'amplitude des coefficients de Fourier est invariante par rapport à la localisation temporelle des signaux, ce qui permet d'étendre la méthode au problème de trouver des séquences similaires ignorant les déplacements (shifts). Avec ce genre de représentations, les séries temporelles sont devenus un objet plus facile à gérer, conduisant à la définition de mesures de similarité efficaces et facilitant l'application des opérations communes de fouille de données.

Dans les travaux [117, 115], nous avons utilisé comme technique de pré-traitement la Transformée en Cosinus Discrète 1D (en anglais Discrete Cosine Transform) (DCT) en réalisant ainsi une réduction de dimensionnalité. Nous proposons de grouper les évolutions ayant un comportement semblable en fréquence, en utilisant la DCT sur l'axe temporel. Cette technique, de transformation du domaine temps dans le domaine fréquence, particulièrement utilisée dans la compression de données, est définie pour transformer des données (corrélées) temporelles en des coefficients (non-corrélés) des formes d'onde à des fréquences progressivement croissantes. Pour des données corrélées, la DCT concentre l'énergie dans les basses fréquences et ainsi, les hautes fréquences peuvent être ignorées sans dégradation significative de qualité.

Les coefficients de DCT sont calculés selon la relation A.4

$$C(u) = \alpha(u) \sum_{i=0}^{I-1} f(i) cos \left[\frac{\pi(2i+1)u}{2I} \right] \tag{A.4}$$

pour u = 0, 1, 2, 3, ..., I-1, le nombre ordinal de forme d'onde

où

$$\alpha(\mathrm{u}) = \begin{cases} \sqrt{\frac{1}{I}} & pour \quad u = 0 \\ \sqrt{\frac{2}{I}} & pour \quad u \neq 0; \end{cases}$$

i = nombre ordinal des images ;

I = nombre total des images de la série ;

$f(i)$ = valeur du pixel.

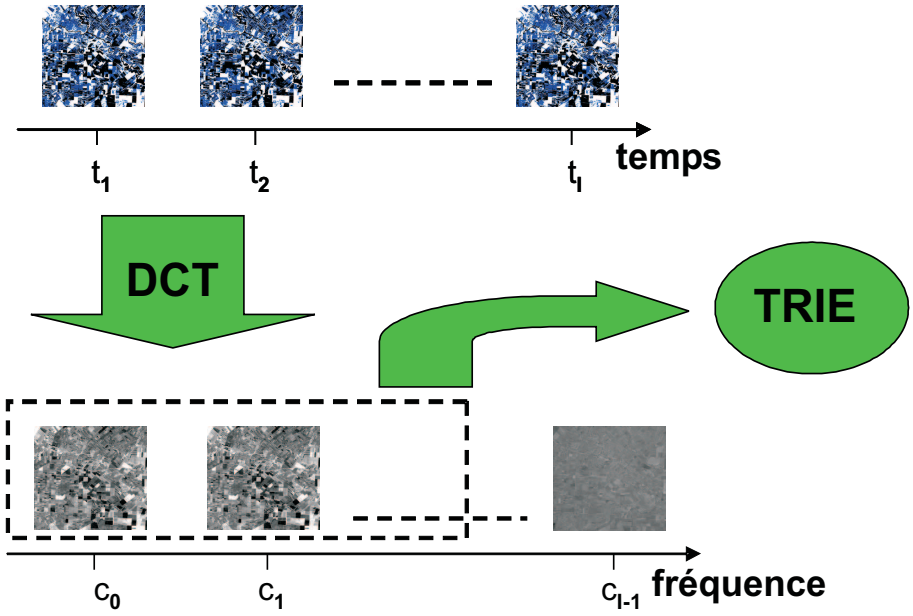
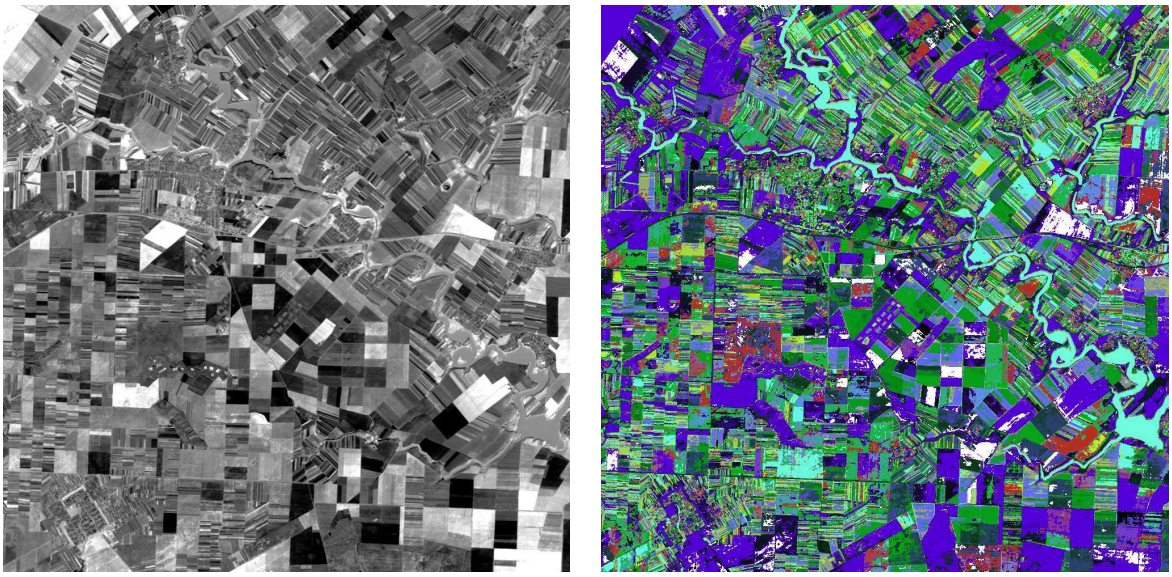


FIG. A.2 – Schéma d'utilisation de la Transformée en Cosinus Discrète



a) b)

FIG. A.3 – a) Image des valeurs du coefficient C_1 de la première forme d'onde obtenue par la Transformation en Cosinus Discrète ; b) Image finale obtenue en utilisant la Transformée Cosinus Discrète sur l'axe temporel (canal B3 ; 253 classes)

Le principe de la méthode est décrit dans la Figure A.2. On fait la transformation des évolutions au niveau du pixel et on obtient une image pour chaque coefficient. Par exemple, la Figure A.3a) présente l'image des valeurs du coefficient C_1 de la première forme d'onde, obtenu par DCT. Dans cette image, nous pouvons remarquer les contours très nets et la bonne homogénéité des zones obtenues. Les images obtenues pour les premiers 5 coefficients sont choisies comme données d'entrée pour l'algorithme de l'arbre de préfixes décrit plus tôt. Pour chaque image de coefficient, les valeurs absolues sont quantifiées en 3 intervalles également peuplés. L'image finale, contenant environ 250 classes, est présentée dans la Figure A.3b). Les frontières sont très bien définies et les régions sont assez homogènes. Si nous comparons l'image ob-

tenue à la vérité terrain, nous observons qu'une même récolte est caractérisée par plusieurs classes. Les cycles phénologiques peuvent être différents dû au fait que les agriculteurs utilisent de différentes variétés d'une même culture et de différentes sortes d'herbicides et d'engrais. Il pourrait également être dû au fait qu'un certain cycle agricole peut être déclenché à des dates différentes. Toutes ces explications semblent pouvoir être valides parce que parties de ces champs appartiennent à un institut pour la recherche dans l'agriculture. De futures démarches sont nécessaires pour grouper thématiquement les classes obtenues.

Annexe B

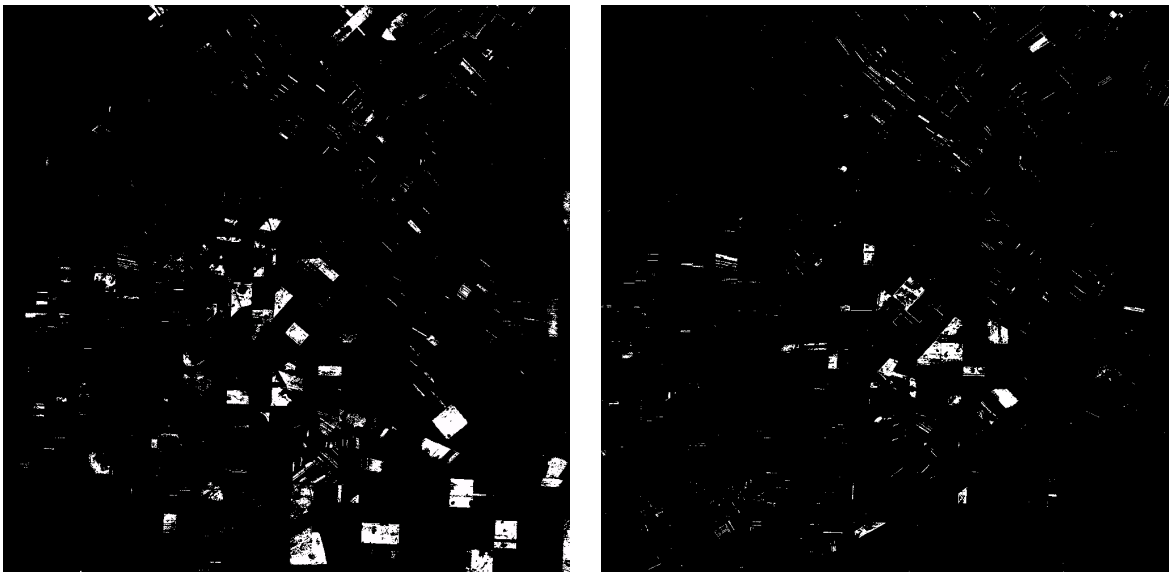
Post-traitements

B.1 Localisation spatiale des motifs séquentiels

Les résultats des tâches d'extraction sont des listes de motifs séquentiels fréquents avec leurs mesures support. De tels résultats sont intéressants, mais ils ne fournissent aucune information sur la localisation spatiale et temporelle des motifs séquentiels.

Dans le but de l'interprétation, de l'évaluation et de la présentation des résultats d'une façon significative, il est nécessaire d'accomplir la localisation spatiale et temporelle des pixels affectés par l'évolution représentée par un motif. À l'aide de ces localisations on obtient des images transformées de celles d'entrée. Leurs pixels detiennent de nouvelles informations sur la localisation spatiale et temporelle d'un motif donné.

Comme post-traitement des MSF extraits, sont réalisées des images avec la localisation spatiale de ces motifs. Dans ces images tous les pixels qui sont concernés par un MSF donné sont allumés, tandis que les autres restent noirs. La Figure B.1 présente la localisation spatiale pour deux MSF incomplets de la STIS ADAM.



a) b)

FIG. B.1 – Localisation spatiale des motifs extraits dans la bande PIR avec σ = 10 000 a) $0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 3 \to 3 \to 3 \to 3 \to 3 \to 3$ b) $2 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 0 \to 3 \to 3$

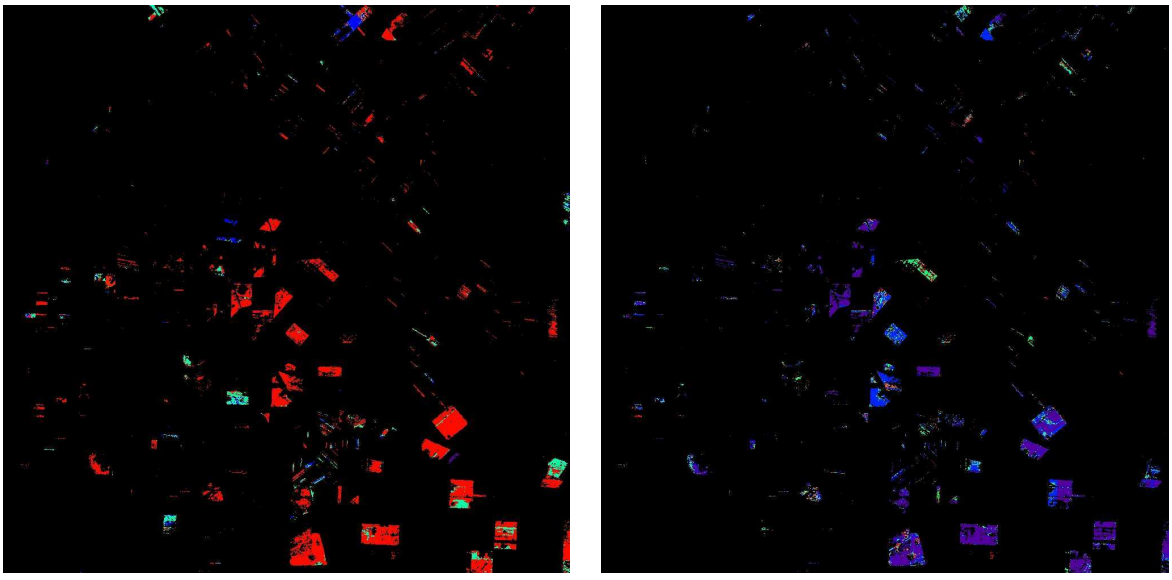
B.2 Localisation temporelle des motifs séquentiels

Puisqu'un motif extrait de longueur incomplète ne contient pas les moments spécifiques de la séquence temporelle dans lequel le motif se produit, il est peut-être utile d'accomplir également, une localisation temporelle. Le traitement suivant est appliqué aux pixels couvert par un motif pour obtenir une nouvelle image de localisation temporelle [123, 115]. Comme dans le cas de la localisation spatiale, les pixels non affectés par l'évolution décrite par le motif donné restent non-éclairés. Pour les pixels affectés par l'évolution représentée par le motif indiqué, la valeur stockée est un nombre de 20 bits ayant le bit i réglé à la valeur 1, où i est le numéro d'ordre de l'image dans laquelle l'évolution donnée a lieu. De cette façon, à chacun des pixels affectés est attribué un nombre à 20 bits. Par exemple, dans le tableau B.1 nous présentons la modalité dans laquelle nous réalisons cette codification pour le motif $0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$. Ainsi le motif de taille 15 est présent aux dates qui ont les numéros d'ordre 1 - 6, 10 - 14, 16 - 18, 20. En mettant en correspondance ces positions avec les puissances de 2 on obtient la codification et l'étiquette indiquée dans le tableau.

Numéro date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Présence motif	x	x	x	x	x	x	-	-	-	x	x	x	x	x	-	x	x	x	-	x
Correspondance	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}	2^{17}	2^{18}	2^{19}
Codification	1	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	1

TAB. B.1 – Codification pour la localisation temporelle du motif avec l'étiquette 1034205

Pour chaque cas de la localisation temporelle du motif donné, nous obtenons une codification différente qui peut être associée à une couleur différente. Des pixels ayant la même évolution aux mêmes dates auront la même couleur.



a) b)

FIG. B.2 – La localisation temporelle des évolutions décrites par le motif $0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$ a) localisation temporelle générale ; b) sur-localisation temporelle des évolutions de l'intervalle de codification 1032703 - 1034205

Les Figures B.1a) et B.1b) montrent qu'il est possible de mettre dans la correspondance un type de culture de la vérité terrain avec une classe extraite par les évolutions des pixels. Dans ces figures, deux situations sont présentées : a) une classe extraite d'évolution des pixels correspond à plusieurs cultures, et b) une classe extraite correspond à seulement une culture.

Afin d'interpréter les résultats nous avons fait une comparaison avec la vraie distribution des cultures dans la période donnée [46]. La Figure B.1a) présente, dans le blanc, la localisation spatiale du motif $0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 0 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3 \rightarrow 3$. Des zones homogènes apparaissent et correspondent aux cultures existantes dans la vérité terrain. Les régions blanches représentées correspondent principalement à la culture de tournesol. Il y a également 3 régions qui correspondent à la culture de pois chiche et à une région correspondant à la culture du soja. Les trois cultures peuvent avoir le même motif mais à différents moments de temps. Le motif a la taille 15 et la STIS a 20 images. Si nous réalisons la discrimination temporelle du motif nous pouvons distinguer les cultures parce qu'elles ont une distribution temporelle différente [123].

Dans la Figure B.2 sont présentées à gauche la localisation temporelle générale d'un motif et à droite un raffinement de cette localisation. Ce raffinement consiste en une sur-discrimination temporelle des évolutions qui dans la Figure B.2a) sont représentées en rouge et qui ont les étiquettes comprises dans l'intervalle 1032703 - 1034205. Ainsi, dans la Figure B.2b) sont visibles surtout 3 couleurs : violet, bleu et vert. La zone avec la couleur verte a l'étiquette 1034205, discutée dans le tableau B.1, et correspond à la culture de soja. De cette façon, la localisation temporelle conduit à la discrimination des cultures agricoles ayant le même motif d'évolution.

Annexe C

Extraction de motifs séquentiels fréquents (premiers résultats)

Le premier concept qui s'est montré utile dans la fouille de données séquentielles est la fréquence, (le support), considérée la première mesure de pertinence des motifs extraits. L'extraction de motifs d'évolution fréquents à partir de STIS a été introduite dans [124, 123, 114]. Les auteurs utilisent ces techniques sur des données optiques et radar, mono ou multi-canal pour l'étude de phénomènes proches et de la surface terrestre.

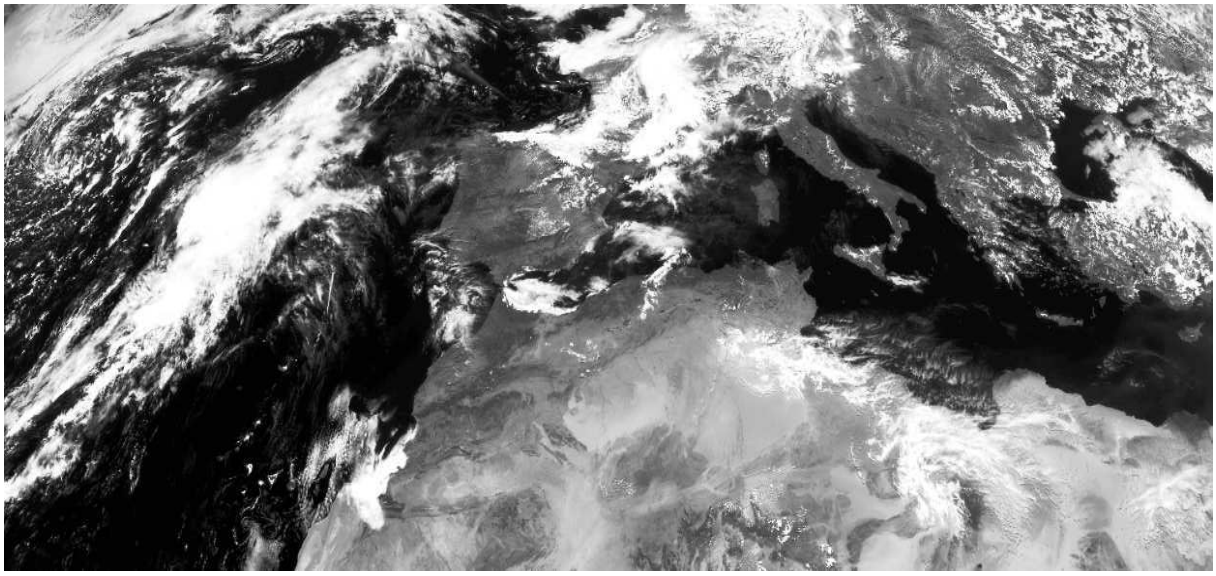
C.1 Extraction de motifs séquentiels de longueur variable

Les données optiques du satellite METEOSAT sont dans le domaine visible ($0,5-0,9\mu m$) et infrarouge thermique ($10,5-12,5\mu m$), et couvrent l'Atlantique du Nord, l'Europe, la région Méditerranéenne et l'Afrique du Nord. Dans le cas des données radar des satellites ERS (Interférométrie Radar à Synthèse d'Ouverture (en anglais Interferometric Synthetic Aperture Radar) (InSAR)), sont utilisées l'amplitude moyenne du signal rétrodiffusé et la cohérence interférométrique, la scène observée étant la région de Mont Blanc, Chamonix, France.

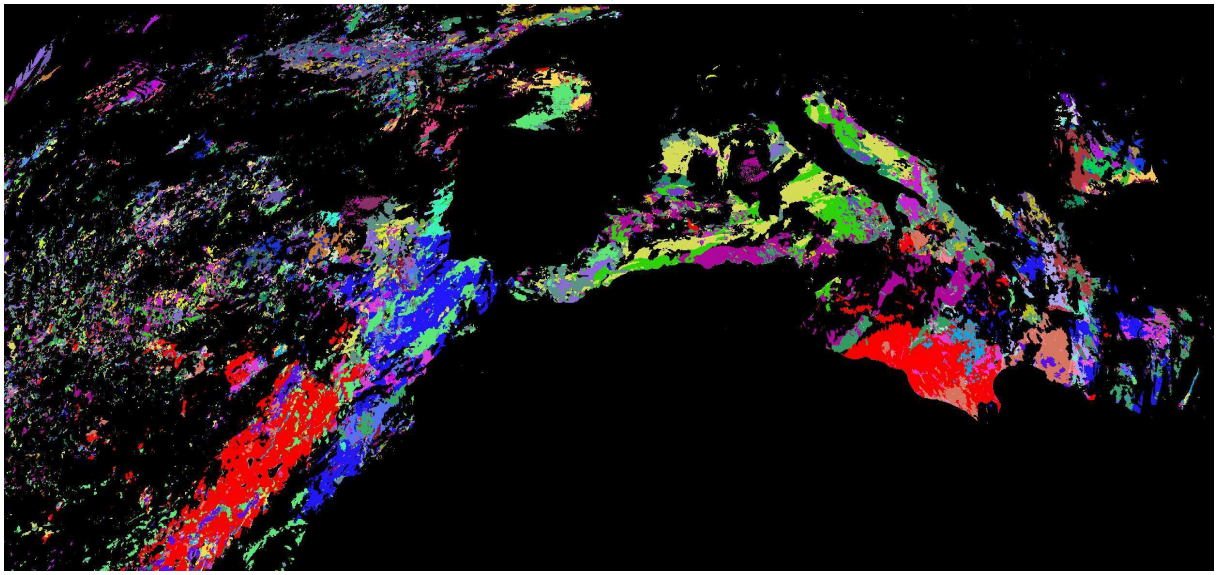
La méthode proposée dans ces travaux pour exhiber les motifs potentiellement utiles est basée sur l'extraction des évolutions des pixels à partir de données de STIS. Les différents types de données (discrétisées par l'expert, une des méthodes décrites dans l'annexe A) ont subi le même traitement pour obtenir de MSF de longueur variable : la fouille de données utilisant le prototype publique [`http://www.cs.rpi.edu`] qui implémente en C++ l'algorithme cSPADE [222], décrit dans la sous-section 3.2.2.

Pour la STIS de 8 images en visible du satellite METEOSAT, un des plus fréquents motifs, $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($supp_{rel} = 17,5\%, s = 4$) est illustré dans la Figure C.1b). Le symbole «0» signifie la présence de l'eau ou de la végétation dense et le symbole «3» indique la présence des nuages très blanches ou de la neige. Dans la Figure C.1a) est présentée une des images d'entrée de la série utilisée. La localisation spatiale et temporelle est montrée dans la Figure C.1b). Le motif est localisé en principal dans les zones maritimes, les continents restent noirs n'étant pas affectés par le motif. La localisation temporelle est réalisée selon la méthode décrite dans l'annexe B ; par exemple, la couleur rouge dénotant la localisation des pixels qui ont l'évolution dans le premier, deuxième, sixième et septième jour.

Un des MSF obtenus à partir de la STIS de 5 paires de données radar des satellites ERS en tandem - amplitude moyenne et cohérence - est le motif multi-canal $17 \rightarrow 17 \rightarrow 17$ ($supp_{rel} =$



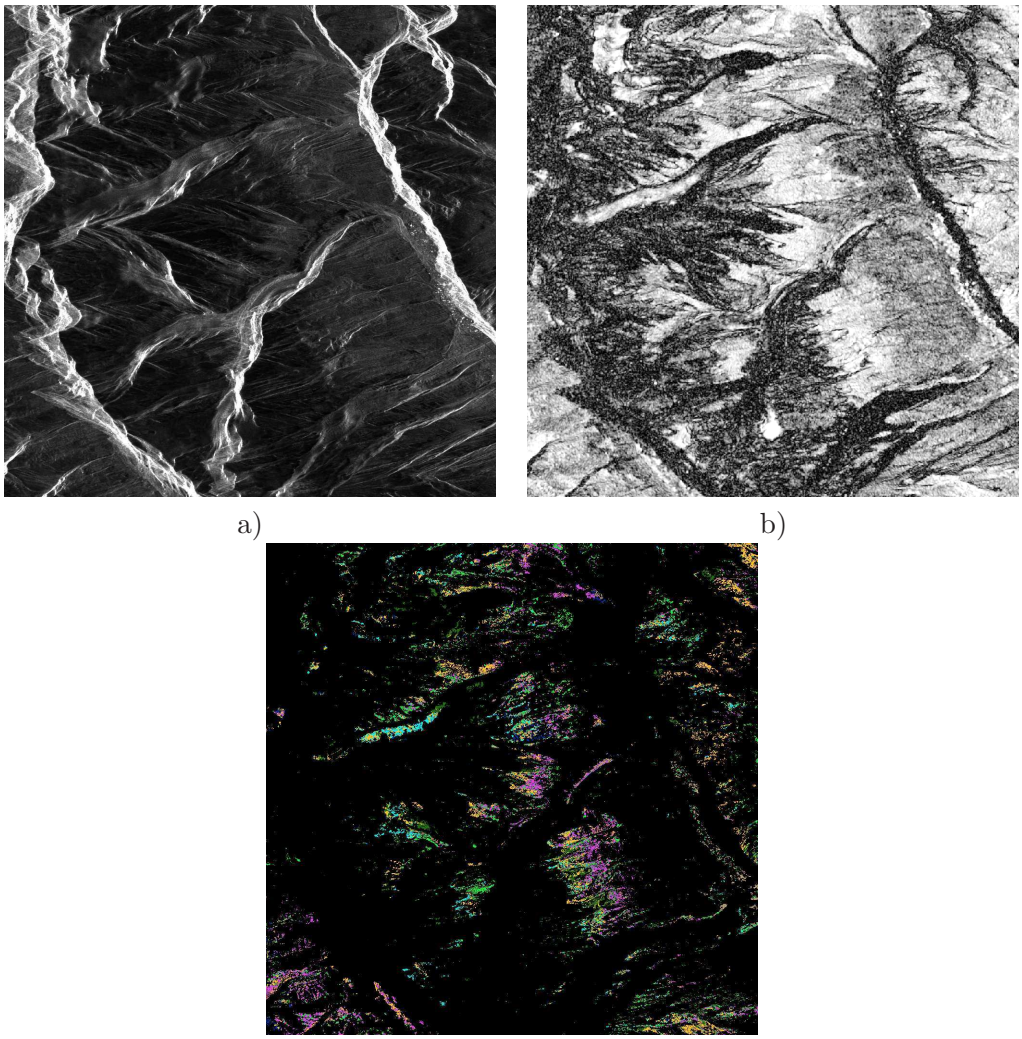
a)



b)

FIG. C.1 – La STIS METEOSAT a) l'image en visible sur la zone observée (13/05/2006) ; la localisation spatiale et temporelle du MSF $0 \rightarrow 0 \rightarrow 3 \rightarrow 0$ ($s = 4, supp_{rel} = 17,5\%$).

7,5%) qui est présenté dans la Figure C.2c). La discrétisation, effectuée par un expert en 4 intervalles pour chaque canal, donne au symbole «1» la signification d'une amplitude faible de rétrodiffusion et au symbole «1» celle d'une cohérence élevée. Par exemple, la région colorée en bleu clair est la part la plus haute des glaciers Argentière et Talèfre et correspond à l'évolution décrite pour les mois Octobre 1996, Mars 1997 et Avril 1997. Cette localisation temporelle révèle un comportement significatif pour des glaciers.



a) b)

c)

FIG. C.2 – Images d'ERS tandem d'octobre 1995 (a) amplitude et (b) la cohérence ; (c) la localisation spatiale et temporelle du MSF $17 \rightarrow 17 \rightarrow 17$ ($s = 4, supp_{rel} = 7,5\%$).

C.2 Extraction de motifs séquentiels de longueur complète - Trie

Pour une caractérisation non redondante de l'évolution au niveau du pixel, il est préférable de tenir compte seulement de motifs séquentiels de longueur maximale, fait qui associe à un pixel une seule évolution.

L'approche de [117, 125, 115] caractérise chaque pixel par une seule évolution afin de les classifier en employant une seule classe et de visualiser toutes les classes dans une seule image.

Une manière efficace pour classifier ces motifs séquentiels est de construire une structure compacte de données telle qu'un arbre de préfixes (trie) comme proposé par de la Briandais [61] et Fredkin [78]. Une telle structure a été employé couramment, par exemple pour améliorer les techniques de télécommunication et de fouille de données (par exemple [217, 178]). Un trie peut garder d'une manière optimisée toutes les évolutions des pixels par le stockage des préfixes qui sont partagés par les motifs séquentiels, seulement une fois.

Il est facile de rechercher ces évolutions parce que chaque chemin de la racine à une feuille

se réfère à une évolution donnée. Considérons l'exemple simple représenté dans la Figure C.3. Dans cet exemple, l'ensemble de données d'entrée est une STIS qui contient 4 images de 9 pixels chacune (p_1 à p_9, dans l'ordre raster). Trois symboles ('rouge', 'bleu' et 'jaune') sont définis pour décrire des valeurs de réflectivité dans toute la STIS. Ces images avaient été acquises aux moments t_1, t_2, t_3 et t_4. Pour chaque pixel, on réalise les séquences temporelles de la Figure C.4 et nous insérons son évolution dans un arbre de préfixes (voir la Figure C.5). Si une évolution existe déjà, c'est-à-dire un chemin complet depuis le noeud racine à une feuille correspondant à cette évolution existe, alors aucune branche ou noeud n'est créé. Respectivement, si il n'y a aucun chemin correspondant, alors les noeuds et les branches appropriés sont créés. Comme on peut observer dans la Figure C.5, à chaque feuille, nous lions un symbole de classe aussi bien que le nombre d'apparitions (dénommé *support*) et les positions de ces apparitions. Par exemple, l'évolution $rouge \rightarrow bleu \rightarrow rouge \rightarrow rouge$ correspond aux pixels p_1 et p_4 et ces pixels seront dans la classe C_1. Ceci correspond à la branche gauche de l'arbre de préfixes. Les premiers deux noeuds (par rapport au noeud racine) de cette branche stockent également le préfixe commun $rouge \rightarrow bleu$ partagé par les classes C_1, C_2 et C_3.

Comme on peut observer, cette structure de données tient compte efficacement des redondances en stockant seulement une fois les préfixes communs. Nous avons décidé d'implémenter cet arbre de préfixes comme proposé dans [61], c'est-à-dire nous plaçons tous les fils d'un noeud dans une liste chaînée qui est elle-même liée à ce noeud (père). Dans Sussenguth [65], on se réfère à cette implémentation comme à un ***arbre doublement chaîné***. Ce genre d'arbre est un arbre binaire car chaque noeud a des liaisons avec maximum deux autres noeuds. Dans un tel arbre, nous assignons de la mémoire seulement pour les noeuds et les feuilles existants par opposition à l'implémentation proposée dans [78]. Dans ce dernier, pour chaque noeud non-feuille, les listes chaînées sont remplacées par des vecteurs dont la taille est égale au nombre de symboles utilisé pour décrire les valeurs de réflectivité du pixel. Chaque cellule est liée alors à un symbole et indique vers le vecteur du fils approprié. Ainsi, un arbre doublement chaîné est clairement plus efficace en parlant de l'espace de stockage tandis que le temps passé pour la sélection du fils approprié d'un noeud n'est plus constant comme il est quand on emploie des vecteurs. En effet, au pire, nous devons parcourir l'entière liste chaînée des fils. Néanmoins, les temps de recherche et ajout se sont avérés proportionnels à $\frac{1}{2}(s+1)log_s C$ où s est la taille moyenne de l'ensemble filiale (nombre moyen de fils pour un père donné) et C est le nombre d'évolutions distinctes contenues dans l'ensemble de données d'entrée. Pour plus de détails, nous référons le lecteur à [65]. C'est-à-dire, une fois que l'arbre est construit, dans le pire des cas, chaque pixel a sa propre évolution, les temps de recherche/ajout sont proportionnels au $\frac{1}{2}(s+1)log_s P$ avec P le nombre de pixels d'une image. Pendant le processus de construction de l'arbre, C et s ne sont pas constants et C augmente toujours sans qu'il dépasse P à la fin du processus. Par conséquent, une limite supérieure inaccessible pour le temps de construction de l'arbre est $P * \frac{1}{2}(s+1)log_s P$, qui est encore correcte même si $P = 1\,000\,000$. Dans notre cas, à chaque feuille, c'est-à-dire à chaque classe d'évolution, nous associons une liste chaînée contenant le nombre d'occurrences dans le premier noeud et les positions des occurrences des évolutions dans les noeuds suivants. Pendant

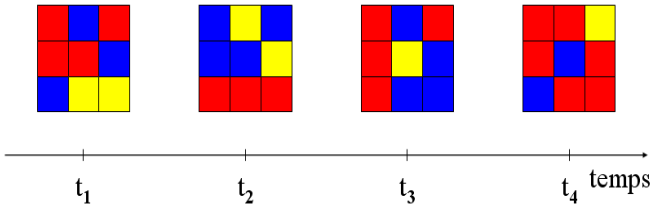


FIG. C.3 – Exemple simple de séquences d'images ($I = 4$, $N = 3$, $P = 9$)

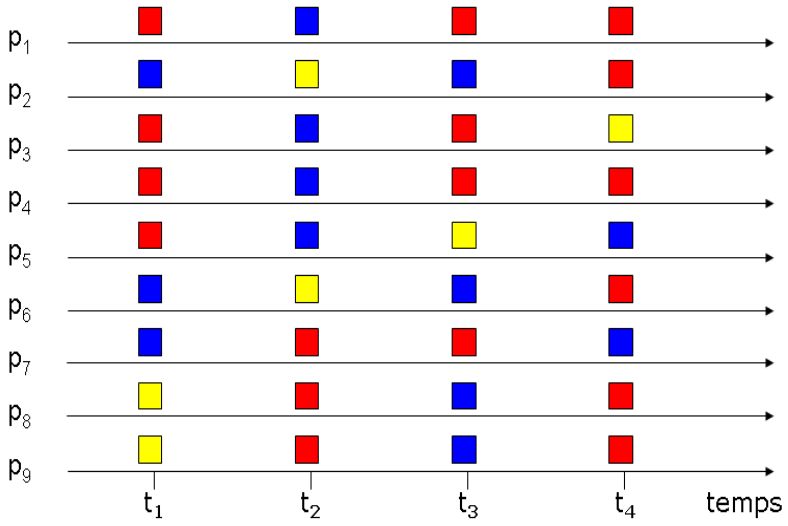


FIG. C.4 – Les séquences temporelles d'évolution des pixels pour la Figure C.3

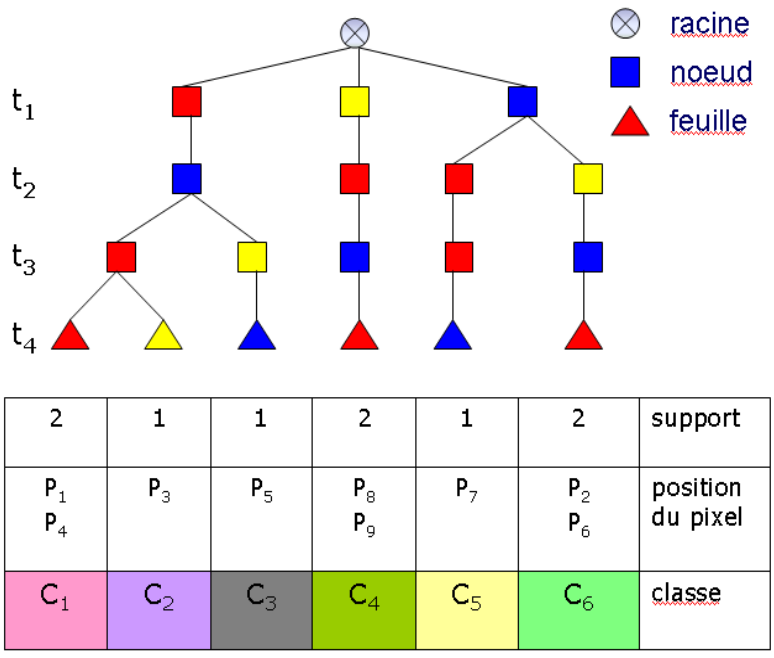


FIG. C.5 – L'arbre de préfixes de la séquence d'images de la Figure C.3

la construction de l'arbre, si une évolution a déjà été stockée pour un pixel donné, et si nous lisons une autre même évolution pour un autre pixel, alors nous mettons à jour le nombre d'occurrences qui est stocké dans le premier noeud et nous insérons, juste après ce noeud, la position des nouvelles occurrences. De cette façon, nous ne devons pas parcourir l'entière liste chaînée pour la mettre à jour. Nous nous référerons au nombre d'occurrences d'une classe d'évolution en tant que *le support d'évolution* ou simplement *support*. De nouveau à notre exemple, dans la Figure C.5, les classes d'évolution C_1, C_4 et C_6 ont un support de 2, signifiant qu'elles ont 2 occurrences chacune. Nous proposons de visualiser toutes les classes d'évolution ou de visualiser les classes d'évolution dont le support appartient à un intervalle défini par l'utilisateur. C'est-à-dire, nous proposons de choisir les classes en tenant compte de la surface qu'elles couvrent, en pixels.

Une autre propriété intéressante est que nous devons parcourir l'ensemble de données seulement une fois pour construire cet arbre. En outre, si nous modifions légèrement cette structure de données, nous pouvons stocker le support des motifs séquentiels aussi bien que les positions des pixels qui sont concernés par ces motifs séquentiels. En conséquence, en construisant un tel arbre de préfixes, nous proposons de classifier les pixels selon leurs évolutions qui sont tracées par des motifs séquentiels et d'obtenir une image d'étiquettes. Dans cette approche, la différence principale avec la méthode présentée dans l'annexe C.1 est que nous pouvons localiser de divers types d'évolutions tout en étant assurés d'avoir une seule évolution par pixel, c'est-à-dire que nous avons un seul type de localisation, celle spatiale.

Pour raffiner cette approche on a réalisé des post-traitements spécifiques comme la régularisation spatiale, la fusion des résultats avec des classes d'évolutions fréquentes sur plusieurs canaux et la fusion des segmentations à l'aide d'une classification non-supervisée des évolutions complètes des pixels.

C.3 Régularisation spatiale

Afin de réduire les effets d'une quantification assez grossière et d'augmenter le degré d'occupation de pixels, nous prenons en considération l'information spatiale et la similitude entre les classes d'évolutions découvertes.

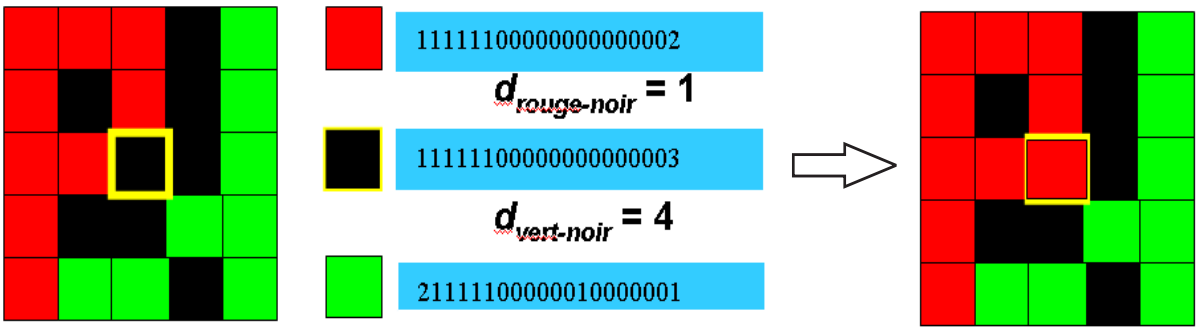


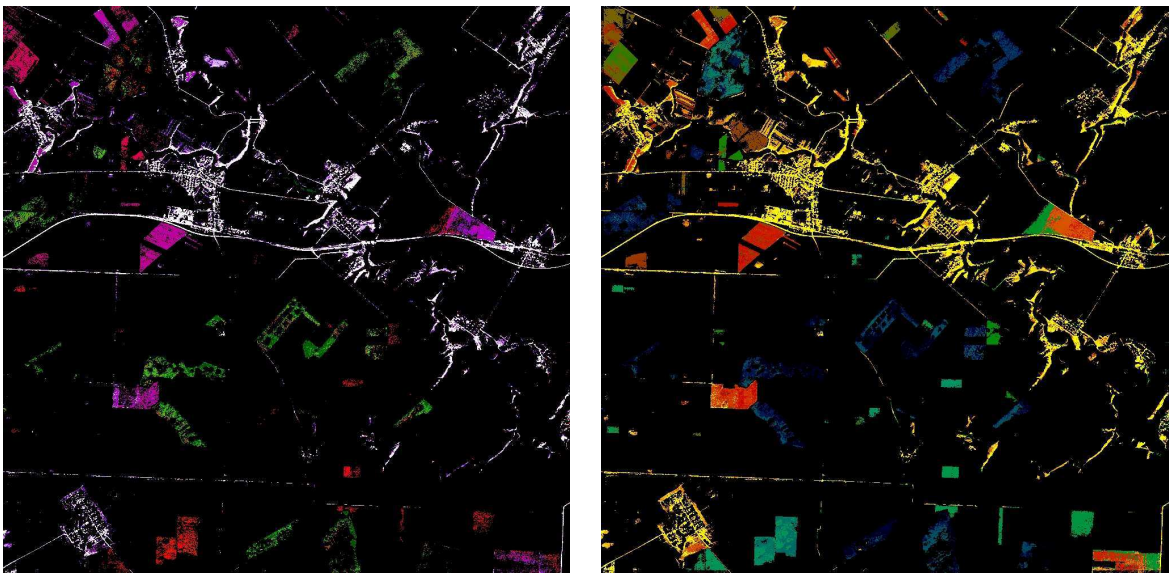
FIG. C.6 – Schéma de la régularisation spatiale

Cette étape de transformation a été dénommée ***régularisation spatiale*** [125, 115]. Elle consiste en l'association des pixels noirs, nonclassifiés, au classes sélectionnées, selon les critères de proximité spatiale et de la similitude des évolutions. Pour chaque pixel noir p_n, qui n'appartient à aucune classe trouvée fréquente (illustrée ici en rouge et jaune), nous centrons une fenêtre de LxL pixels. Dans la Figure C.6 est présenté le schéma de ce traitement pour $L =$

5. Dans cette fenêtre, nous recherchons la classe d'évolution visualisée la plus similaire à la classe du pixel central, pixel mis en évidence dans la diagramme gauche de la Figure C.6. Les séquences d'évolutions sont traitées comme des vecteurs et la mesure de dissimilitude employée est la distance de Manhattan (distance de Minkowski de premier ordre). Pour les vecteurs 1 et 2, cette distance est :

$$d_{2-1} = \sum_{i=1}^{N} |x_{2i} - x_{1i}| \tag{C.1}$$

Nous calculons la distance entre les vecteurs définis par les séquences d'évolution correspondant au pixel p_n et les pixels colorés contenus dans la fenêtre et nous identifions la distance minimale. Si cette distance est plus petite qu'un seuil, w, défini par l'utilisateur, le pixel p_n est assigné à la classe qui correspond aux pixels colorés les plus semblables. Par exemple, dans la Figure C.6, pour $w = 3$, le pixel central est attribué à la classe de pixels colorés en rouge, parce que $d_{rouge-noir}$, qui est la distance minimale dans la fenêtre, est inférieure au seuil w choisi. S'il y a plusieurs classes qui ont la distance minimale, la classe la plus peuplée dans la fenêtre est préférée. S'il y a plusieurs classes avec la même population parmi les candidats, le prochain critère est la fréquence dans l'image globale. Si la condition du seuil défini par l'utilisateur n'est pas remplie ou il n'y a pas des pixels colorés dans la fenêtre, le pixel p_n reste noir. Le procédé de la régularisation spatiale peut être également employé dans le cas d'une image avec des tonalités de gris.

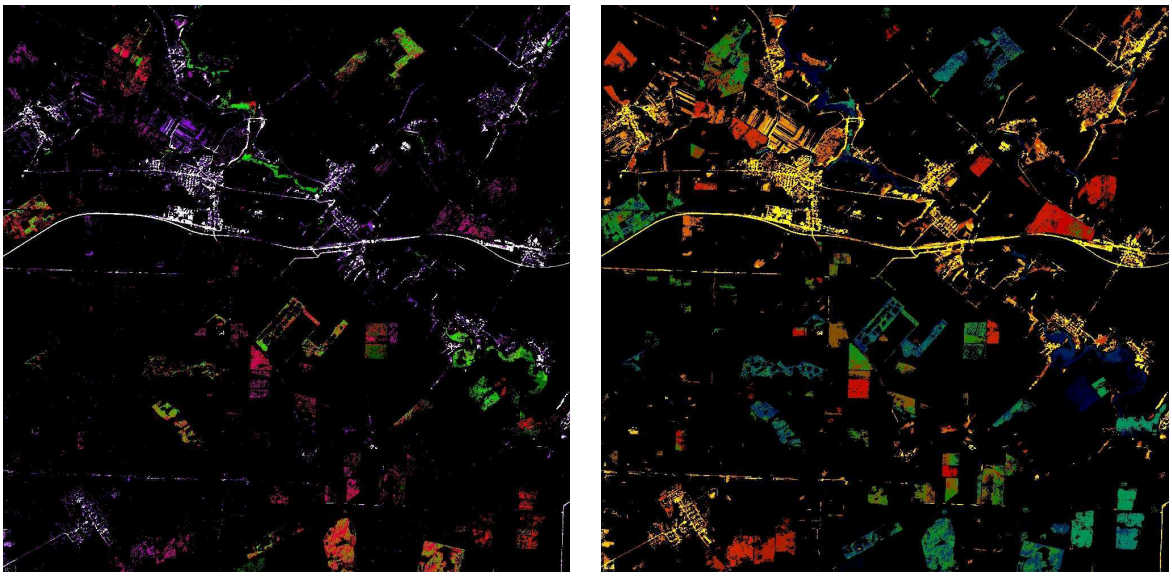


a) b)

FIG. C.7 – a) Localisation de 166 classes d'évolution extraites à partir de la bande B1 de la STIS ADAM avec $\sigma = 100$ b) la même localisation après régularisation spatiale avec $L = 5$ et $w = 3$.

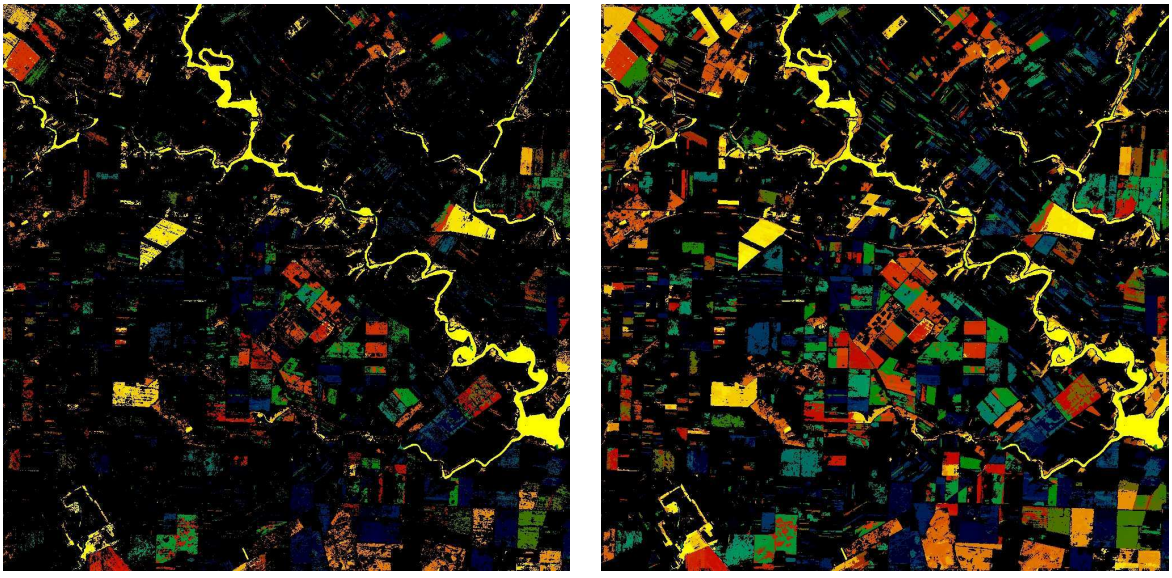
Les régularisations spatiales des images des Figures C.7a, C.8a et C.9a sont présentées dans les Figures C.7b, C.8b et C.9b. Nous pouvons noter l'augmentation de la homogénéité des classes représentées aussi bien que l'augmentation du nombre de pixels colorés. Dans le cas de la Figure C.9, pour la bande PIR, le degré d'occupation est plus que doublé par la régularisation spatiale (de 167.776 à 339.379 pixels).

En tenant compte du caractère agricole de la scène observée, nous avons aussi utilisé une bande synthétique B4. Cette bande est établie en calculant l'*IVDN* en utilisant les bandes B2



a) b)

FIG. C.8 – a) Localisation de 205 classes d'évolution extraites à partir de la bande B2 de la STIS ADAM avec $\sigma = 100$ b) la même localisation après régularisation spatiale avec $L = 5$ et $w = 3$.



a) b)

FIG. C.9 – a) Localisation de 561 classes d'évolution extraites à partir de la bande B3 (PIR) de la STIS ADAM avec $\sigma = 100$ b) la même localisation après régularisation spatiale avec $L = 5$ et $w = 3$.

et B3. Dans la Figure C.10, on peut voir le résultat de la régularisation spatiale pour les classes fréquentes avec support absolu de 100 pour la bande synthétique B4. Les zones avec spécifique agricole et les forêts sont mieux caractérisées que les zones des villages et des routes.

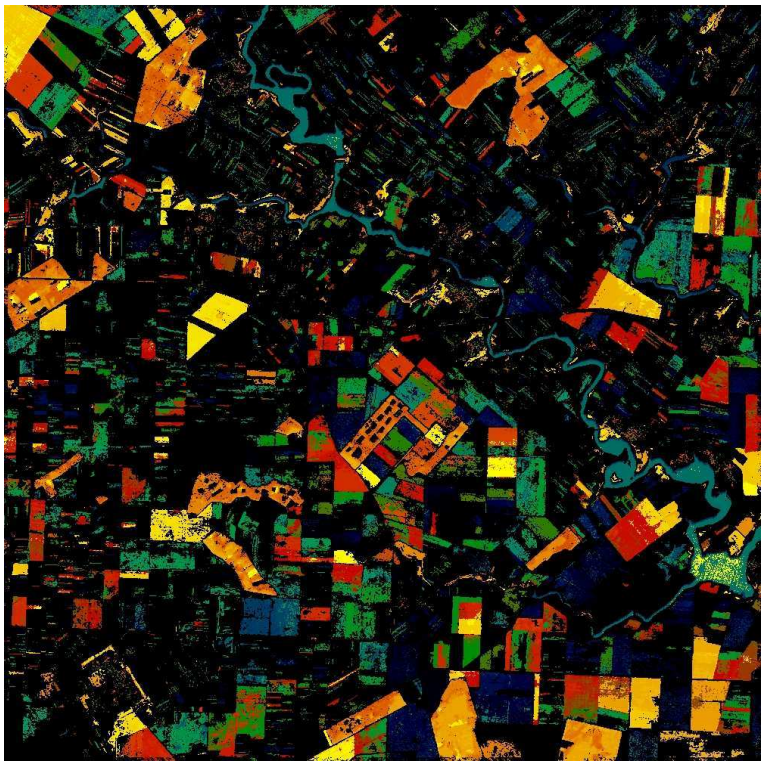


FIG. C.10 – Localisation de 538 classes d'évolution extraites à partir de la bande B4 (IVDN) de la STIS ADAM avec $\sigma = 100$ après régularisation spatiale avec $L = 5$ et $w = 3$

C.4 Fusion des résultats avec des classes d'évolutions fréquentes sur plusieurs canaux

Une autre méthode d'augmentation du degré d'occupation du sol dans le cas des images avec des classes d'évolutions fréquentes, mais qui conduit aussi à la croissance du nombre de classes, est la fusion des résultats sur plusieurs canaux. Comme exposé dans [117], chaque canal de la STIS utilisée apporte sa propre spécificité concernant l'information thématique de la scène. Afin d'obtenir une caractérisation plus riche de la scène observée et l'augmentation de l'occupation de pixels, nous proposons de combiner nos résultats en superposant les images des classes d'évolution régularisées spatialement obtenues en utilisant le même seuil σ pour les bandes B1, B2, B3 et la bande synthétique B4. Si un pixel peut être lié à plusieurs classes d'évolutions, c'est-à-dire il y a au moins deux bandes pour lesquelles une classe d'évolution peut être assignée au pixel, nous assignons d'abord le pixel à une classe B4. S'il n'y a aucune classe B4, alors nous recherchons d'autres classes en considérant les bandes dans l'ordre suivant : B3, B2, B1. Un résultat de ce type de fusion est présenté dans la Figure C.11. L'occupation de pixels monte jusqu'à 51.23% tandis que l'homogénéité des régions augmente aussi en comparaison avec les anciennes images [125, 115]. Comme la fusion accorde la priorité au canal B4 de l'IVDN, le contenu végétal de la scène est bien caractérisé.

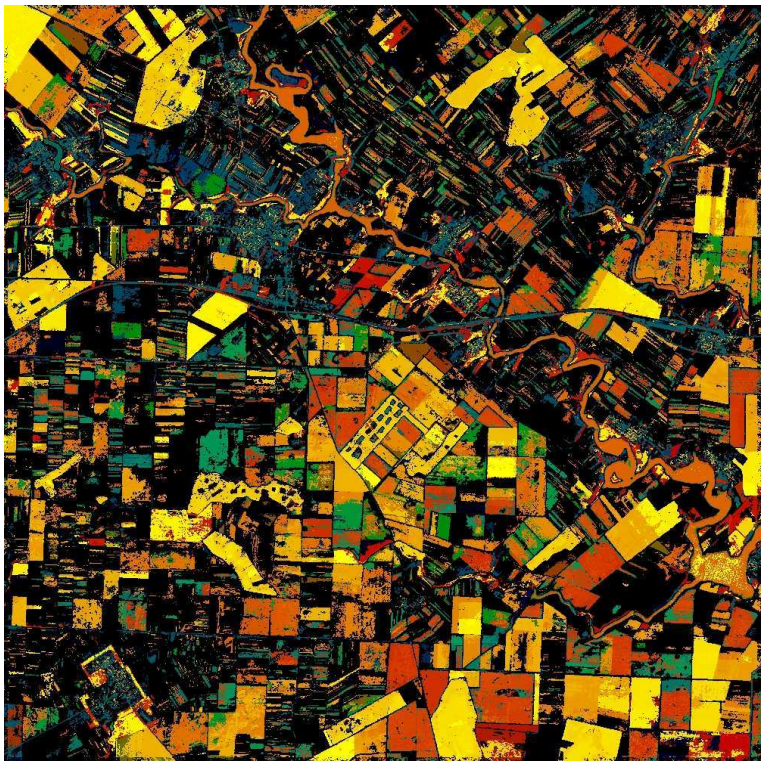
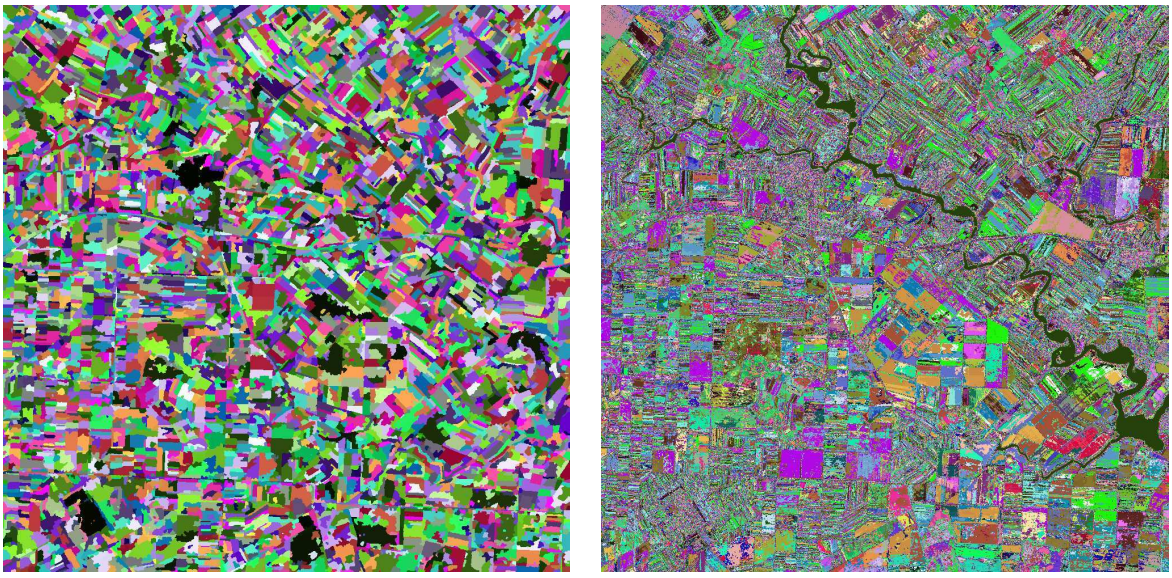


FIG. C.11 – L'image finale des évolutions avec 1467 classes après la fusion de résultats avec des classes d'évolutions fréquentes sur plusieurs canaux ($\sigma = 100$, $L = 5$, et $w = 3$)

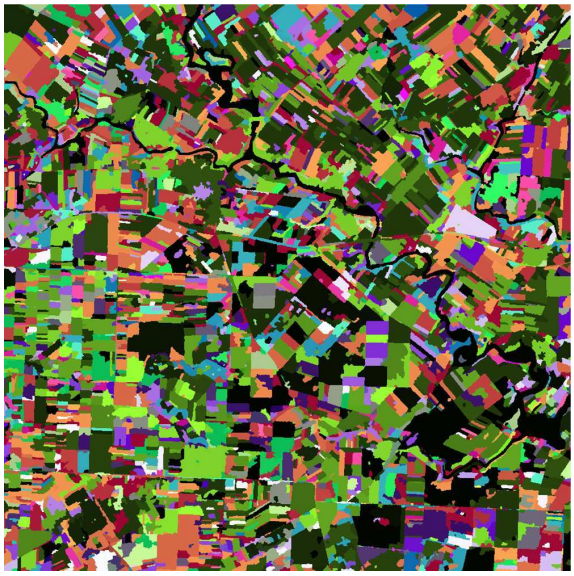
C.5 Fusion des segmentations à l'aide d'une classification non-supervisée des évolutions complètes des pixels

La classification obtenue par l'extraction des évolutions au niveau du pixel implique une composante temporelle essentielle pour une STIS. Ce type d'information peut être combiné avec n'importe quel autre résultat obtenu par l'analyse des images d'une STIS. Par exemple, nos cartes d'évolutions de la STIS, basées sur l'analyse au niveau du pixel, peuvent raffiner des segmentations réalisées sur chaque image de la série. Ainsi, les segmentations sont enrichies par des informations concernant les évolutions temporelles. Ces segmentations sont obtenues par une extension de la méthode Longueur de Description Minimale (en anglais Minimum Description Length) (MDL) [147]. Une telle segmentation est visualisée dans la Figure C.12 a). L'image de classes d'évolution, présentée dans la Figure C.12b), est obtenue à l'aide de l'arbre de préfixes à partir des quantifications en 4 classes réalisées avec l'algorithme K-moyennes.

La fusion de la série de segmentations et de l'image de classification est réalisée par la méthode du vote majoritaire : à chaque région de la segmentation nous assignons la classe d'évolution la plus représentée parmi les pixels de cette région [146, 115]. Un résultat d'une telle sorte de fusion est présentée dans la Figure C.12c). En comparaison avec la segmentation initiale, on observe la décroissance du nombre de segments et que des régions avec signification thématique commencent d'être aperçues. Par exemple, la rivière Mostiştea devient clairement visible dans l'image finale et quelques champs agricoles se définissent plus précisément.



a) b)



c)

FIG. C.12 – Fusion d'une segmentation avec une classification d'évolutions basées sur le pixel - images d'entrée a) segmentation d'une image de la STIS obtenue avec la méthode MDL ; b) image de classes d'évolution de la STIS c) Image obtenue par la fusion d'une segmentation avec une classification d'évolutions basées sur le pixel

Annexe D

La phénoménologie et la phénologie de la scène de la STIS ADAM

Dans cette annexe les mécanismes qui influencent les propriétés spectrales des principaux objets de la scène sont présentés. La nature physique et physiologique de ces objets peut être déduite par leur signature spectrale et l'évolution temporelle de leur réponse spectrale peut permettre leur identification.

D.1 La végétation

La végétation verte présente des minimums de réflectance dans le visible, où l'absorption domine. Pour le proche infrarouge les niveaux des réflectance et transmission sont hauts et l'absorption est très réduite (Figure 6.1).

Visible
La photosynthèse est le processus qui assure la synthèse des composés organiques nécessaires pour l'entretien de la vie et la croissance des plantes. Le processus est déclenché par l'absorption de la part visible du rayonnement solaire au niveau des chloroplastes. Ces cellules contiennent des pigments (chlorophylle et carotène) capables d'absorber dans les domaines spectraux bleu et rouge. Ainsi, le vert, étant transmis ou réfléchi, donne un maximum dans la caractéristique spectrale de réflectance et la couleur de la plante vive. Vers la finalisation de la vie fonctionnelle de la plante, les tissus des feuilles et des autres organes se détériorent. C'est la sénescence, quand la chlorophylle se détruit et la plante commence à réfléchir également dans le domaine spectral du rouge.

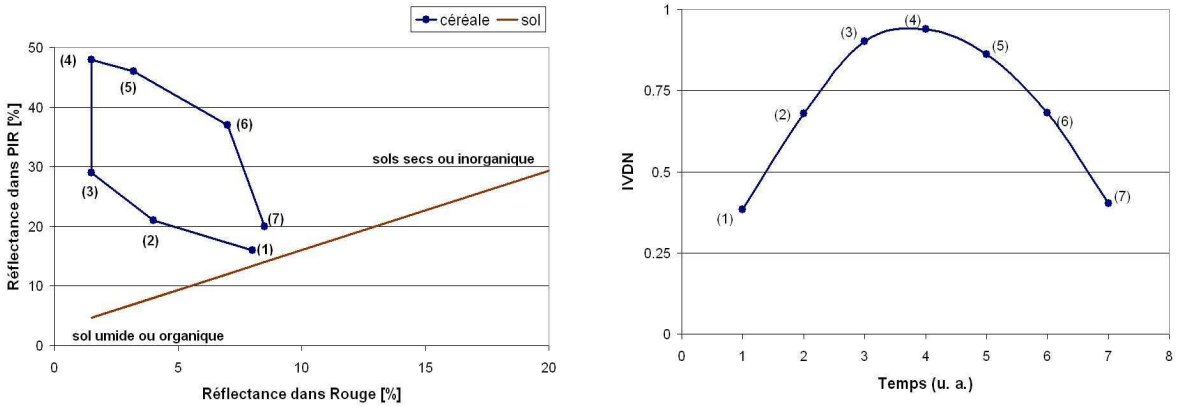
Proche infrarouge
L'énergie correspondante au domaine proche infrarouge ($0,8 - 1,1 \mu m$) n'est pas suffisante pour déclencher les réactions photochimiques du cycle photosynthétique et les pigments des chloroplastes sont transparents dans cette région spectrale. Typiquement l'absorption est très faible ($\approx$ 5%) et la réponse spectrale est dominée par la transmission ($\approx$ 40%) et par la réflexion ($\approx$ 55%). La proportion entre ces valeurs varie avec l'espèce de plante, et elle est contrôlée par la structure interne des feuilles. Pendant la sénescence, la réflectance dans le PIR change. Au début, elle peut croître mais dans les stades avancés elle diminue drastiquement.

Les valeurs typiques énoncées au dessus sont pour une feuille. Le rayonnement transmis par

une feuille peut rentrer dans d'autres feuilles où la proportion réfléchi / transmis se conserve. Pour un modèle avec ce rapport 50%/50% après un nombre infini d'incidences la valeur maximale de la réflectance, ρ^{∞}, peut atteindre la valeur de 84% [166]

D.2 Le sol nu

Les propriétés spectrales du sol sont plus simples puisque les phénomènes impliqués sont la réflexion et l'absorption, la transmission étant nulle. Dans la région spectrale étudiée, pour les sols secs, la réflectance croît continuellement avec la longueur d'onde du rayonnement incident. Cette tendance peut être cachée par l'effet produit par la variation du contenu d'eau qui diminue proportionnellement avec la réflectivité. On peut voir l'influence de l'humidité du sol dans la Figure D.1a) où la croissance du contenu d'eau approche le point représentatif du sol de l'origine du graphique. La forme de base est valable pour la majorité des sols, c'est seulement la magnitude qui varie [49]. Les principales variables qui peuvent influencer les propriétés spectrales sont la texture, l'humidité, le contenu de matière organique et le contenu d'oxydes de fer. En dépit de ce fait, il y a une relation linéaire entre les réflectances des deux longueurs d'onde et en général elle reste constante spatialement et temporellement. Ainsi, si un sol a une grande réflectance dans le visible, il aura un grande réflectance aussi dans le proche infrarouge, et vice-versa (Figure D.1).



a) b)

FIG. D.1 – a) Les caractéristiques spectrales d'un cycle végétal et du sol [21] et b) Le cycle végétal d'une céréale transposé en IVDN.

D.3 L'eau

D'une manière similaire au sol et à la végétation, l'eau présente des variations de la réflectance avec la longueur d'onde du rayonnement incident. Les changements en réponse spectrale dépendent de modifications produites dans les états physiques, chimiques et biologiques de l'eau. De la même manière comme les propriétés spectrales diffèrent entre un sol sec et un humide, entre une végétation verte et une sénescente, la même situation peut être observée entre une eau claire et une trouble, entre l'eau d'un océan profond et celle d'un lac petit. Les propriétés spectrales de l'eau claire montrent pour le domaine visible une excellente transmission et une insignifiante absorption, et pour le proche infrarouge une forte absorption et une faible transmission. L'eau se comporte comme un réflecteur spéculaire quand la géométrie émetteur - surface cible - capteur le permet. Dans la Figure 6.1 est présentée la caractéristique spectrale de l'eau d'une rivière trouble qui a des sédiments et des suspensions, le cas de notre scène.

D.4 Considérations temporelles - la phénologie

Toutes les plantes ont deux processus majeurs pendant leur cycle de vie : la croissance (modifications quantitatives comme le volume, la forme ou les fonctions) et la différentiation (modifications qualitatives comme le passage de la phase végétative à la phase reproductive).

Pour décrire les transformations saisonnières des plantes, le terme phénologie est utilisé fréquemment. L'évolution typique de la réflectance d'une plante au cours de son développement [21] est décrite par la courbe bleue de la Figure D.1a) qui présente les changements spectraux dans les domaines rouge et proche infrarouge pour une céréale. Lorsque les plantes germent (point 1), la réflectance dans le rouge commence à diminuer en raison de l'absorption par la chlorophylle, et la réflectance dans le proche infrarouge commence à augmenter lentement car la végétation réfléchit dans le proche infrarouge. La position précise du point initial, dans l'espace VIS - PIR, dépend de la réflectance du sol. La figure présente aussi une dépendance approximative de la réflectance d'un sol entre les deux situations extrêmes : sol sec ou inorganique et sol très humide ou de nature organique. Dans l'évolution décrite par le segment (2) - (3), plus la couverture végétale augmente, moins on voit le sol, et donc plus la réflectance dans le proche infrarouge augmente. La réflectance dans le rouge continue de diminuer aussi jusqu'à atteindre un minimum lié au fait que les nouvelles feuilles sous la canopée sont totalement occultées et leur contribution à la réflexion est nulle. Ceci correspond au point (3), quand on atteint la réflectance ρ^{∞} pour le rouge. En revanche, après cela, la réflectance dans l'infrarouge continue de croître avec le nombre de feuilles jusqu'au point (4), où l'on atteint la réflectance ρ^{∞} pour le proche infrarouge. Lors de la sénescence, la chlorophylle se décompose, ce qui se traduit par un accroissement de la réflectance dans le rouge, et le changement de l'orientation des cellules des feuilles a pour conséquence une baisse de la réflectance dans l'infrarouge ((4) - (5)). De suite, la réflectance dans le rouge croît et la diminution de celle dans le proche infrarouge s'accélère en raison de modifications suivantes dans la structure des feuilles ((5) - (6)). Finalement, la culture est récoltée, laissant le sol et les résidus post-récoltes (point (7)).

Le comportement qui est décrit correspond à un comportement moyen. Les valeurs des différents points ainsi que les intervalles temporels sont spécifiques à chaque plante et dépendent aussi de la zone géographique et du mode de culture.

L'évolution temporelle spectrale du cycle végétal de la Figure D.1a) est convertie dans l'évolution de l'IVDN présentée dans la Figure D.1b).

L'influence de l'humidité du sol nu peut être comprise avec la Figure D.2 qui transforme approximativement la variation de courbe du sol de la Figure D.1a) en variation de l'IVDN. Un sol très humide peut altérer l'information phénologique.

Si on relève une évolution d'un pixel correspondant à une culture d'hiver et à une culture d'été on peut voir l'influence du degré d'humidité du sol provoquée par de précipitations plus accentuées. Dans la Figure D.3 qui présentent les évolutions en IVDN des pixels couverts par les cultures de maïs et respectivement de blé, la distribution décadaire des précipitations, [19], est superposée. De cette manière on peut expliquer quelques variations inhabituelles et brusques des courbes phénologiques et observer le degré différent d'influence dans ces deux cas de cultures. L'échelle temporelle est exprimée en jours commençant par 31 octobre 2000. Pour le maïs, les précipitations plus consistantes de mars 2001 trouvent le sol nu et ainsi on peut expliquer les sauts de la courbe d'IVDN comme l'influence de l'humidité excessive. En revanche, au mois de mars, les terrains avec du blé sont presque couverts par la végétation et les sauts dans la courbe d'IVDN ne sont pas aussi évidents.

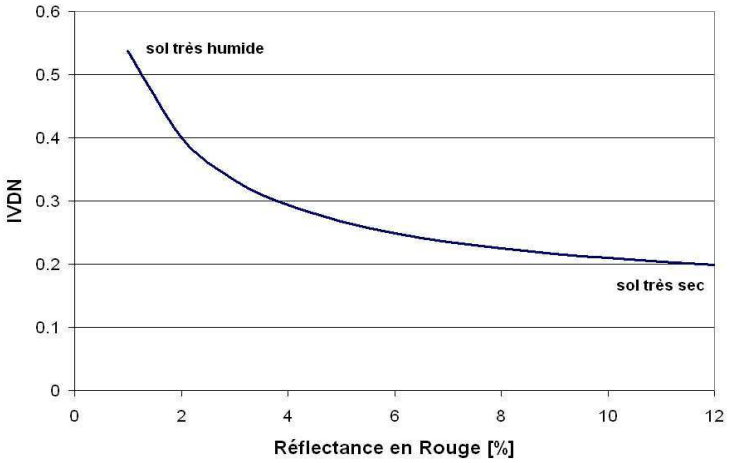
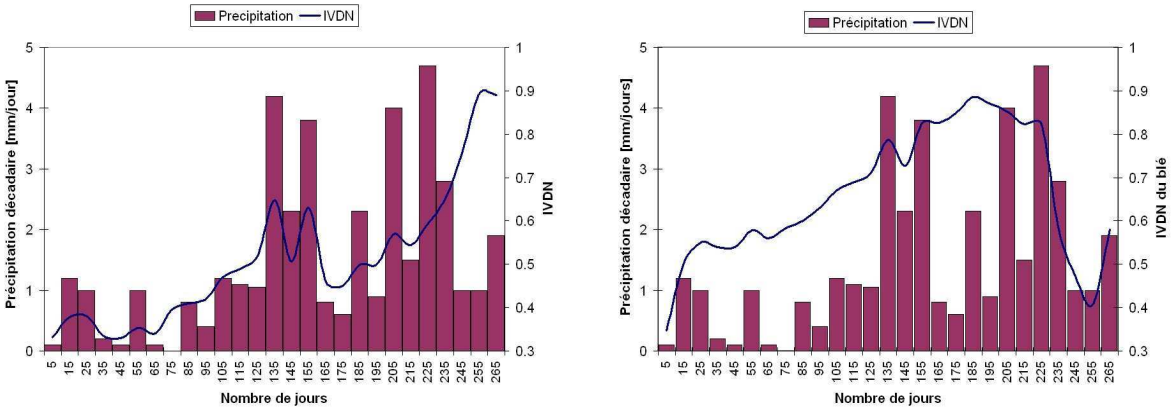


FIG. D.2 – La dépendance approximative de l'IVDN du sol avec l'humidité



a) b)

FIG. D.3 – La courbe phénologique obtenue pour a) le maïs et b) le blé en comparaison avec la précipitation décadaire de la période octobre 2000 - juillet 2001.

Bibliographie

[1] http://www.spotimage.com/web/fr/148-les-satellites-spot.php, Les satellites SPOT, online. 80

[2] http://earth.esa.int/ers/, ESA Missions - ERS, online. 128

[3] http://envisat.esa.int/, ESA Missions - ENVISAT, online. 128

[4] http://www.ccrs.nrcan.gc.ca/resource/tutor/polarim/index_e.php, Tutorial : Radar Polarimetry, Canada Centre for Remote Sensing, Natural Resources Canada, online. 136

[5] http://www.radarsat2.info/about/mission.asp, Mission RADARSAT-2, online. 136

[6] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the ACM SIGMOD International Conference on Management Data*, pages 207–216, Washington, DC, USA, 1993. ACM Press. 15, 34, 40

[7] R. Agrawal, T. Imielinski, and A. Swami. Sequentional Pattern Mining Using Bitmap Representation. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, Alberta, Canada, 2002. 41

[8] R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zaït. Querying shapes of histories. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB '95)*, pages 502–514, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 16

[9] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994. 34, 45

[10] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Proc. of the 11th International Conference on Data Engineering (ICDE'95)*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press. 2, 27, 34, 38, 39, 40, 43, 46, 86, 88

[11] F. Amelung, D. L. Galloway, J. W. Bell, H. A. Zebker, and P. J. Laczniak. Sensing the ups and downs of Las Vegas : InSAR reveals structural control of land subsidence and aquifer-system deformation. *Geology*, 27(6) :483–486, 1999. 130, 134

[12] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M.-J. Kraak, A. MacEachren, and S. Wrobel. Geovisual analytics for spatial decision support : Setting the research agenda. *International Journal of Geographical Information Science*, 21(8) :839–857, 2007. 14

[13] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization : an analytical review. *Journal of Visual Languages and Computing*, 14(6) :503–541, December 2003. Special Issue on Visual Data Mining. 14

[14] C. Antunes. *Pattern Mining over Nominal Event Sequences using Constraint Relaxations*. PhD thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisbon, Portugal, January 2005. 45

[15] P. Aplin, P. Atkinson, and P. Curran. Fine spatial resolution satellite sensors for the next decade. *International Journal of Remote Sensing*, 18 :1387–1381, 1997. 21

[16] P. Aplin and G. Smith. Advances in object-based image classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII, 2008. Part B7. 22

[17] R. Athauda, M. Tissera, and C. Fernando. Data Mining Applications : Promise and Challenges. In J. Ponce and A. Karahoca, editors, *Data Mining and Knowledge Discovery in Real Life Applications*, pages 201–214. InTech, Vienna, Austria, 2009. 14

[18] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential Pattern Mining Using Bitmap Representation. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 429–435, Edmonton, Alberta, Canada, July 2002. 39, 41, 47

[19] F. Baret, R. Vintilă, C. Lazăr, N. Rochdi, L. Prévot, J. C. Favard, H. Deboissezon, C. Lauvernet, P. Voicu, E. Petcu, G. Petcu, J.-P. Denux, O. Marloie, C. Radnea, D. Turnea, C. Simota, V. Poenaru, F. Cabot, and P. Henry. Preliminary results of the ADAM Project : investigating high temporal revisit frequency at high spatial satellite resolution for crop monitoring. In A. Canarache and R. Enache, editors, *Proc. Int. Conf. Soils under Global Change - a Challenge for the 21st Century*, volume 1, pages 63–78, 2002. 80, 173

[20] R. J. Bayardo. The Hows, Whys, and Whens of Constraints in Itemset and Rule Discovery. In *Proceedings of the European Workshop on Inductive Databases and Constraint Based Mining*, pages 1–13, 2004. 48, 70

[21] A. S. Belward. Spectral Characteristics of Vegetation, Soil and Water in the Visible, Near-Infrared Wavelength. In A. S. Belward and C. R. Valenzuela, editors, *Remote Sensing and Geographical Information Systems for Resource Management in Developing Countries*, pages 31–53. Kluwer Academic Publishers, 1991. v, ix, 82, 172, 173

[22] M. Bertolotto, S. Di Martino, F. Ferrucci, and M.-T. Kechadi. Visualization system for collaborative spatio-temporal data mining. *Journal of Geographical Information Science*, 21(7), 2007. 14

[23] W.-M. Boerner, H. Mott, E. Lunenburg, C. Livingstone, B. Brisco, R. J. Brown, and J. S. Patterson. Polarimetry in Radar Remote Sensing : Basic and Applied Concepts - Chapter 5. In F. M. Henderson and A. J. Lewis, editors, *Principles and Applications of Imaging Radar*, volume 2 of *Manual of Remote Sensing*, pages 271–358. John Wiley ans Sons, New York, USA, third edition, 1998. 135

[24] F. Bonchi and F. Giannotti. Pushing Constraints to Detect Local Patterns. In K. Morik, J.-F. Boulicaut, and A. Siebes, editors, *Local Pattern Detection, International Seminar, Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, volume 3539 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2005. 27, 29, 44

[25] F. Bonchi and C. Lucchese. Pushing tougher constraints in frequent pattern mining. In T. B. Ho, D. W.-L. Cheung, and H. Liu, editors, *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD'05, Hanoi, Vietnam, May 18-20, 2005*, volume 3518 of *Lecture Notes in Computer Science*, pages 114–124. Springer, 2005. 48

[26] F. Bonchi and C. Lucchese. Extending the state-of-the-art of constraint-based pattern discovery. *Data Knowl. Eng.*, 60 :377–399, February 2007. 44

[27] S. Bontemps, P. Bogaert, N. Titeux, and P. Defourny. An object-based change detection method accounting for temporal dependences in time series with medium to coarse spatial resolution. *Remote Sensing of Environment*, 112 :3181–3191, 2008. 26

[28] S. Boriah, V. Kumar, M. Steinbach, C. Potter, and S. Klooster. Land cover change detection : a case study. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, pages 857–865, New York, NY, USA, 2008. ACM. 26

[29] J.-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00)*, volume 1805 of *Lecture Notes in Artificial Intelligence*, pages 62–73, Kyoto, Japan, 2000. Springer-Verlag. 28

[30] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by mean of free-sets. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '00)*, volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 75–85, Lyon, France, September 2000. Springer-Verlag. 28

[31] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery Journal*, 7(1) :5–22, 2003. 28

[32] W. Boulila, I. R. Farah, K. S. Ettabaa, B. Solaiman, and H. B. Ghézala. Spatio-temporal modeling for knowledge discovery in satellite image databases. In *Proceedingsof the 7th French Information Retrieval Conference, COnférence en Recherche d'Infomations et Applications - CORIA 2010, Sousse, Tunisia, March 18-20, 2010*, pages 35–49. Centre de Publication Universitaire, 2010. 32

[33] L. Bruzzone and D. Fernández Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3) :1171–1182, May 2000. 26

[34] B. Buttenfield, M. Gahegan, H. Miller, and M. Yuan. Geospatial data mining and knowledge discovery. UCGIS white paper on Emergent Research Themes, 2001. 18

[35] A. Bykowski and C. Rigotti. A condensed representation to find frequent patterns. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS'01)*, pages 267–273, Santa Barbara, CA, USA, May 2001. ACM Press. 28

[36] A. Bykowski and C. Rigotti. DBC : A condensed representation of frequent patterns for efficient mining. *Information Systems*, 28(8) :949–977, 2003. 28

[37] Y. Cai, D. Clutter, G. Pape, J. Han, M. Welge, and L. Auvil. MAIDS : Mining Alarming Incidents from Data Streams. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2004. 43

[38] T. Calders and B. Goethals. Mining all non derivable frequent itemsets. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, volume 2431 of *Lecture Notes in Artificial Intelligence*, pages 74–85, Helsinki, Finland, August 2002. Springer-Verlag. 28

[39] T. Calders and B. Goethals. Minimal k-free representations of frequent sets. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, volume 2838 of *Lecture Notes in Artificial Intelligence*, pages 71–82, Cavtat-Dubrovnik, Croatia, September 2003. Springer-Verlag. 28

[40] T. Calders, C. Rigotti, and J.-F. Boulicaut. A survey on condensed representation for frequent sets. In J.-F. Boulicaut, L. Raedt, and H. Mannila, editors, *Proceedings of Constraint-Based Mining and Inductive Databases 2004*, volume 3848 of *Lecture Notes in Computer Science*, pages 64–80. Springer-Verlag, 2005. 28

[41] J. B. Campbell. *Introduction to Remote Sensing*. The Guilford Press, New York, USA, third edition, 2002. 20

[42] H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05)*, pages 82–89, Washington, DC, USA, 2005. IEEE Computer Society. 31

[43] H. Cao, N. Mamoulis, and D. W. Cheung. Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering*, 19(4) :453–467, 2007. 31

[44] H. Carrao, P. Gonsalves, and M. Caetano. A nonlinear harmonic model for fitting satellite image time series : Analysis and prediction of land cover dynamics. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4) :1919–1930, 2010. 29

[45] O. Cavalié, M.-P. Doin, C. Lasserre, and P. Briole. Ground motion measurement in the Lake Mead area, Nevada, by differential synthetic aperture radar interferometry time series analysis : Probing the lithosphere rheological structure. *Journal of Geophysical Research*, 112, 2007. 128, 129

[46] Centre National d'Etudes Spatiales. Database for the Data Assimilation for Agro-Modeling (ADAM) project. online. `http://kalideos.cnes.fr/index.php?id=accueil-adam`. 3, 80, 157

[47] S. Cloude and E. Pottier. A review of target decomposition theorems in radar polarimetry. *IEEE Transactions on Geoscience and Remote Sensing*, 34(2) :498–518, March 1996. 137

[48] S. Cloude and E. Pottier. An entropy based classification scheme for land applications of polarimetric SAR. *IEEE Transactions on Geoscience and Remote Sensing*, 35(1) :68–78, January 1997. viii, 137, 138

[49] H. R. Condit. The spectral reflectance of american soils. *Photogrammetric Engineering*, 36(9) :955–966, 1970. 172

[50] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin. Digital change detection methods in ecosystem monitoring : a review. 25(9) :1565–1596, May 2004. 26

[51] T. Cox and M. Cox, editors. *Multidimensional Scaling*. Monographs on statistics and applied probability. Chapman & Hall/CRC, 2001. 12

[52] P. Crapper. An estimate of the number of boundary cells in a mapped landscape coded to grid cells. *Photogrammetric Engineering and Remote Sensing*, 50 :1497–1503, 1984. 20

[53] B. Crémilleux and J.-F. Boulicaut. Simplest rules characterizing classes generated by delta-free sets. In *Proceedings of the International Conference on Knowledge Based Systems and Applied Artificial Intelligence*, pages 33–46, Cambridge, UK, 2002. Springer-Verlag. 27

[54] B. Crémilleux and A. Soulet. Discovering knowledge from local patterns with global constraints. In *From Local Patterns to Global Models (LeGo-09), ECMLPKDD'09 Workshop, Bled, Slovenia, 7 September 2009*. 28

[55] E. P. Crist and R. C. Cicone. A Physically Based Transformation of Thematic Mapper Data : the TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, 22 :256–263, 1984. 82

[56] M. Datcu, H. Daschiel, A. Pelizzari, A. G. M. Quartulli, A. Colapicchioni, M. Pastori, K. Seidel, P. Marchetti, and S. D'Elia. Information mining in remote sensing image archives : System concepts. *IEEE Transactions on Geoscience and Remote Sensing*, 41 :2923–2936, 2003. 30

[57] M. Datcu, K. Seidel, and M. Walessa. Spatial Information Retrieval from Remote-Sensing Images. Part I - Information Theoretical Perspectives. *IEEE Transactions on Geoscience and Remote Sensing*, 36 :1431–1445, 1998. 30

[58] S. M. de Jong, E. J. Pebesma, and F. D. van der Meer. Spatial variability, mapping methods, image analysis and pixels. In S. M. de Jong and F. D. van der Meer, editors, *Remote Sensing Image Analysis : Including the Spatial Domain*, chapter 2. Springer, The Netherlands, 2006. 1, 2

[59] S. M. de Jong and F. D. van der Meer, editors. *Remote Sensing Image Analysis : Including the Spatial Domain*. Springer, The Netherlands, 2006. 2

[60] S. M. de Jong, F. D. van der Meer, and J. G. P. W. Clevers. Basics of remote sensing. In S. M. de Jong and F. D. van der Meer, editors, *Remote Sensing Image Analysis : Including the Spatial Domain*, chapter 1. Springer, The Netherlands, 2006. 53, 77

[61] R. De La Briandais. File searching using variable length keys. In *Proc. of the Western Joint Computer Conference*, pages 295–298, New York, 1959. 161, 162

[62] L. de Raedt and A. Zimmermann. Constraint-Based Pattern Set Mining. In *Proc. of the 7th SIAM International Conference on Data Mining*, Minneapolis, Minnesota, USA, April 2007. 48

[63] Deliverable of the EFIDIR project [79]. SPATPAM : a SPAtio-TemPorAl Mining tool. online, July 2009. http ://www.efidir.fr. 64

[64] G. Dong and J. Pei, editors. *Sequence Data Mining*, volume 33 of *Advances in Database Systems*. Springer, 2007. 2, 8, 34, 45, 46, 47, 48

[65] J. Edward H. Sussenguth. Use of tree structures for processing files. *Commun. ACM*, 6(5) :272–279, 1963. 162

[66] M. Egenhofer and J. Sharma. Topological relations between regions in R2 and Z2. In *Proceedings of the 3rd International Symposium on Advances in Spatial Databases (SSD'93), Singapore*, volume 692 of *Lecture Notes in Computer Science*, pages 316–331. Springer-Verlag, 1993. 16

[67] M. Ester, H.-P. Kriegel, and J. Sander. Spatial data mining : A database approach. In *Proceedings of the 5th International Symposium on Advances in Spatial Databases (SSD'97), Berlin, Germany*, pages 47–66, London, UK, 1997. Springer-Verlag. 17

[68] W. J. Ewens and G. R. Grant, editors. *Statistical methods in bioinformatics : An introduction*. Springer-Verlag, 2001. 15

[69] U. Fayyad and G. Grinstein. Introduction. In *Information Visualization in Data Mining and Knowledge Discovery*, pages 1–17. Morgan Kaufmann, Los Altos, CA, USA, 2001. 14

[70] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery : An overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Ulthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. MIT Press, Cambridge, MA, USA, 1996. 1, 17

[71] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge Discovery and Data Mining : Towards a Unifying Framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD*, pages 82–88, 1996. 11, 27

[72] P. Fisher. The pixel : A snare and a delusion. *International Journal of Remote Sensing*, 18 :679–685, 1997. 21

[73] P. Fisher, P. Laube, M. Kreveld, and S. Imfeld. Finding REMO - Detecting Relative Motion Patterns in Geospatial Lifelines. In *Developments in Spatial Data Handling*, pages 201–215. Springer Berlin Heidelberg, 2005. 31

[74] R. Fisher, K. Dawson-Howe, A. Fitzgibbon, C. Robertson, and E. Trucco, editors. *Dictionary of Computer Vision and Image Processing*. John Wiley and Sons, New York, USA, 2005. 57

[75] G. Foody. Fully fuzzy supervised classification of land cover from remotely sensed imagery with an artificial neural network. *Neural Computing and Applications*, 5 :238–247, 1997. 20

[76] G. Foody. The Continuum of Classification Fuzziness in Thematic Mapping. *Photogrammetric Engineering and Remote Sensing*, 65(4) :443–451, 1999. 21

[77] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. Knowledge discovery in databases : an overview. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge in Discovery in Databases*, pages 1–27. Menlo Park : AAAI Press, 1991. 10

[78] E. Fredkin. Trie memory. *Commun. ACM*, 3(9), 1960. 161, 162

[79] French national research (ANR) project. Extraction and Fusion of Information for measuring ground displacements with Radar Imagery (EFIDIR) project. online. `http://www.efidir.fr`. 3, 64, 126, 179

[80] J. Fürnkranz. From local to global patterns : Evaluation issues in rule learning algorithms. In *Local Pattern Detection, International Seminar, Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, volume 3539 of *Lecture Notes in Computer Science*, pages 20–38. Springer, 2005. 27

[81] L. Gallucio, O. Michel, and P. Comon. Unsupervised clustering on multi-components datasets : Applications on images and astrophysics data. In *16th European Signal Processing Conference EUSIPCO-2008*, pages 25–29, Lausanne, Switzerland, August 2008. 25

[82] P. Gançarski and C. Wemmert. Collaborative multi-strategy classification : application to per-pixel analysis of images. In *Proc. of the 6th International Workshop on Multimedia Data Mining : mining integrated media and complex data*, pages 15–22, 2005. 24

[83] M. Garofalakis, R. Rastogi, and K. Shim. SPIRIT : Sequential Pattern Mining with Regular Expression Constraints. In *Proc. of the 25th International Conference on Very Large Databases (VLDB'99)*, pages 223–234, Edinburgh, United Kingdom, September 1999. 39, 40, 45, 46

[84] G. Gianella, J. Han, J. Pei, X. Yan, and P. Yu. Mining Frequent Patterns in Data Streams at Multiple Time Granularities. In K. S. H. Kargupta, A. Joshi and Y. Yesha, editors, *Next Generation Data Mining Chapter 3*, 2003. 43

[85] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In P. Berkhin, R. Caruana, and X. Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 330–339. ACM, 2007. 32

[86] B. Goethals. Frequent set mining. In *The Data Mining and Knowledge Discovery Handbook*, chapter 17, pages 377–397. Springer-Verlag, 2005. 28

[87] S. Griguolo. Classification on Sets of Remotely-Sensed Images : a Vegetation Monitoring Model. In E. Binaghi, P. Brivio, and A. Rampini, editors, *Soft Computing in Remote Sensing Data Analysis*, pages 235–244. World Scientific, Singapore, 1996. 21

[88] J. Gudmundsson, M. J. van Kreveld, and B. Speckmann. Efficient detection of motion patterns in spatio-temporal data sets. In D. Pfoser, I. F. Cruz, and M. Ronthaler, editors, *Proceedings of the 12th ACM International Workshop on Geographic Information Systems, ACM-GIS 2004, November 12-13, 2004, Washington, DC, USA*, pages 250–257. ACM, 2004. 31

[89] L. Gueguen. *Extraction d'information et compression conjointes des séries temporelles d'images satellitaires.* PhD thesis, Télécom Paris, École Nationale Supérieure des Télécommunications, Paris, France, Octobre 2007. 29, 30

[90] L. Gueguen and M. Datcu. Image time-series data mining based on the information-bottleneck principle. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4) :827–838, 2007. 30

[91] L. Gueguen and M. Datcu. A similarity metric for retrieval of compressed objects : Application for mining satellite image time series. *IEEE Transactions on Knowledge and Data Engineering*, 20 :562–575, April 2008. 31

[92] M. Guérif, D. Courault, and N. Brisson. Assimilation des données de télédétection dans les modèles de fonctionnement des cultures. In *Actes de l'École - Chercheurs INRA en bioclimatologie, Le Croisic, 1996*, volume 2, pages 169–191, Le Croisic, France, March 1997. 80

[93] J. Han. Data Mining. In J. Urban and P. Dasgupta, editors, *Encyclopedia of Distributed Computing*. Kluwer Academic Publishers, 1999. 13

[94] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining : current status and future directions. *Data Mining and Knowledge Discovery*, 15(1) :55–86, 2007. 14

[95] J. Han and M. Kamber, editors. *Data Mining : Concepts and Techniques*. The Morgan Kaufmann Series in Data Management. Morgan Kaufmann Publishers, first edition, 2000. 11, 14, 15

[96] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data)*, volume 29 of *SIGMOD Rec.*, pages 1–12, Dallas, Texas, USA, 2000. ACM. 34

[97] D. Hand. Pattern Detection and Discovery. In D. Hand, N. Adams, and R. Bolton, editors, *Pattern Detection and Discovery*, volume 2447 of *Lecture Notes in Computer Science*, pages 161–173. Springer Berlin / Heidelberg, 2002. 27, 28

[98] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001. 14, 15

[99] J. D. Hand. Data mining : Statistics and more ? *The American Statistician*, 52(2) :112–118, 1998. 14

[100] G. Hay and G. Castilla. Object-based image analysis : Strengths, weaknesses, opportunities and threats (SWOT). In *Proceedings of the 1st International Symposium on Object-based Image Analysis (OBIA 2006), Salzburg University, Austria, July 4-5, 2006*, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS), 2006. 22

[101] P. Héas. *Apprentissage Bayésien de Structures Spatio-Temporelle : application à la fouille visuelle de séries temporelles d'images de satellites*. PhD thesis, École Nationale Supérieure de l'Aéronautique et de l'Espace, France, Avril 2005. 29

[102] P. Héas and M. Datcu. Modeling trajectory of dynamic clusters in image time-series for spatio-temporal reasoning. *IEEE Transactions on Geoscience and Remote Sensing*, 43(7) :1635–1647, 2005. 30

[103] F. M. Henderson and A. J. Lewis. Introduction - Chapter 1. In F. M. Henderson and A. J. Lewis, editors, *Principles and Applications of Imaging Radar*, volume 2 of *Manual of Remote Sensing*, pages 1–6. John Wiley ans Sons, New York, USA, third edition, 1998. 126

[104] J. Hipp and U. Güntzer. Is pushing constraints deeply into the mining algorithms really what we want ? : an alternative approach for association rule mining. *ACM SIGKDD Explorations*, 4(1) :50–55, 2002. 45

[105] R. Honda and O. Konishi. Temporal rule discovery for time-series satellite images and integration with RDB. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '01)*, pages 204–215, London, UK, 2001. Springer-Verlag. 22

[106] R. Honda, S. Wang, T. Kikuchi, and O. Konishi. Mining of moving objects from time-series images and its application to satellite weather imagery. *J. Intell. Inf. Syst.*, 19 :79–93, July 2002. 22

[107] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4) :433–448, April 2008. 18, 32

[108] A. R. Huete and R. D. Jackson. The Suitability of Spectral Indices for Evaluating Vegetation Characteristics on Arid Rangelands. *Remote Sensing of Environment*, 23 :213–232, 1987. 82

[109] J. Inglada, J.-C. Favard, H. Yesou, S. Clandillon, and C. Bestault. Lava flow mapping during the Nyiragongo January, 2002 eruption over the city of Goma (D.R. Congo) in the frame of the international charter space and major disasters. In *Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'03)*, volume 3, pages 1540–1542, 2003. 25

[110] J. Han and J. Pei and B. Mortazavi-Asl and Q. Chen and U. Dayal and M. Hsu. FreeSpan : frequent pattern-projected sequential pattern mining. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, USA, 2000. 39, 41

[111] B. Jeudy. *Optimisation des requêtes inductives : Application à l'extraction sous contraintes de règles d'association.* PhD thesis, L'Institut National des Sciences Appliquées de Lyon, Lyon, France, Décembre 2002. 2, 10

[112] W. Johnston. Model visualization. In *Information Visualization in Data Mining and Knowledge Discovery*, pages 223–227. Morgan Kaufmann, Los Altos, CA, USA, 2001. 14

[113] A. Julea. Transformation et simulation d'images en géométrie Radar à Synthèse d'Ouverture, 2005. projet fin d'études Universitatea Politehnica Bucuresti - Université de Savoie. 126

[114] A. Julea. Sequential patterns in satellite imagery. Master's thesis, Universitatea Politehnica Bucuresti, 2007. 8, 159

[115] A. Julea. Extraction des évolutions à partir de séries temporelles d'images satellitaires, Octobre 2008. Rapport LISTIC no.07-08. 8, 152, 156, 161, 164, 167, 168

[116] A. Julea, F. Ledo, N. Méger, E. Trouvé, P. Bolon, C. Rigotti, R. Fallourd, J. Nicolas, G. Vasile, M. Gay, O. Harant, L. Ferro-Famil, and F. Lodge. PolSAR RADARSAT-2 Satellite Image Time Series Mining over the Chamonix Mont-Blanc Test Site. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2011)*, pages 1191–1194, Vancouver, Canada, 2011. 59, 78, 88, 126, 136, 140, 145, 150, 152

[117] A. Julea, N. Méger, and P. Bolon. On mining pixel based evolution classes in satellite image time series. In *Proc. of the 5th Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008)*, ESRIN - Frascati, Italy, March 2008. 21, 27, 35, 78, 130, 150, 152, 161, 167

[118] A. Julea, N. Méger, P. Bolon, and V. Lăzărescu. Spatiotemporal mining of evolutions in Satellite Image Time Series. *Scientific Bulletin of University Politehnica of Bucharest, Series C : Electrical Engineering and Computer Science*, 2011. submitted. 59, 64, 70, 98

[119] A. Julea, N. Méger, P. Bolon, C. Rigotti, M.-P. Doin, C. Lasserre, E. Trouvé, and V. Lăzărescu. Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns. *IEEE Transactions on Geoscience and Remote Sensing*, 49(4) :1417–1430, 2011. 58, 59, 67, 78, 88, 105, 126, 130, 131, 133, 135, 144, 145, 150

[120] A. Julea, N. Méger, C. Rigotti, M. P. Doin, C. Lasserre, E. Trouvé, P. Bolon, and V. Lăzărescu. Extraction of frequent grouped sequential patterns from satellite image time series. In *Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010)*, volume 5, pages 3434–3437, Honolulu, Hawaii, USA, 2010. 57, 58, 59, 67, 78, 88, 126, 130, 145, 150

[121] A. Julea, N. Méger, C. Rigotti, E. Trouvé, P. Bolon, and V. Lăzărescu. Mining Pixels Evolutions in Satellite Image Time Series for Agricultural Monitoring. In *Advances in Data Mining. Applications and Theoretical Aspects - Proceedings of the 11th Industrial Conference on Data Mining (ICDM 2011)*, volume 6870/2011 of *Lecture Notes in Computer Science*, pages 189–203, New York, USA, 2011. DOI :10.1109/Multi-temp.2011.6005067.10. 59, 78, 88, 130, 144, 145, 150

[122] A. Julea, N. Méger, C. Rigotti, E. Trouvé, R. Jolivet, and P. Bolon. Efficient Spatiotemporal Mining of Satellite Image Time Series for Agricultural Monitoring. *Transactions on Machine Learning and Data Mining*, 4(2), 2011. ibai publishing, Germany. 59, 78, 88, 130, 144, 145, 150

[123] A. Julea, N. Méger, and E. Trouvé. On mining METEOSAT and ERS multitemporal images. In *Proc. of the 4th Conference on Image Information Mining for Security and Intelligence (ESA-EUSC 2006)*, Madrid, Spain, November 2006. 8, 21, 27, 35, 41, 49, 86, 143, 150, 156, 157, 159

[124] A. Julea, N. Méger, and E. Trouvé. Sequential patterns extraction in multitemporal satellite images. In *10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'06), Practical Data Mining Workshop : Applications, Experiences and Challenges*, pages 94–97, Berlin, Germany, September 2006. 8, 21, 27, 35, 41, 49, 86, 143, 150, 159

[125] A. Julea, N. Méger, E. Trouvé, and P. Bolon. On extracting evolutions from satellite image time series. In *Proc. of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2008)*, volume 5, pages 228–231, Boston, MA, USA, 2008. 8, 21, 27, 35, 78, 130, 150, 161, 164, 167

[126] A. Julea, N. Méger, E. Trouvé, P. Bolon, C. Rigotti, R. Fallourd, J. Nicolas, G. Vasile, M. Gay, O. Harant, and L. Ferro-Famil. Spatio-temporal mining of PolSAR satellite image time series. In *ESA Living Planet Symposium*, Bergen, Norway, 2010. 59, 78, 88, 126, 136, 140, 145, 150, 152

[127] A. Julea, I. Petillot, G. Vasile, E. Trouvé, V. Buzuloiu, and D. Haşegan. Slant Range Rectification Of Georeferenced Information For SAR Data Analysis In Mountainous Regions. In *Proceedings of the 1st International Summer School on Optoelectronic Techniques for Environmental Monitoring and Risk Assessment*, pages 253–258, Baia Mare, Romania, 2006. 126

[128] A. Julea, G. Vasile, I. Pétillot, E. Trouvé, M. Gay, J.-M. Nicolas, and P. Bolon. Simulation of SAR Images and Radar Coding of Georeferenced Information for Temperate Glacier Monitoring. In *Proceedings of the International Conference on Optimization of Electrical and Electronic Equipment*, volume IV, pages 175–180, Braşov, Romania, 2006. 126

[129] T. Kanungo, B. Dom, W. Niblack, and D. Steele. A Fast Algorithm for MDL-Based Multiband Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, June 21-23, 1994*, pages 609–616, 1994. 28

[130] M.-T. Kechadi, M. Bertolotto, F. Ferrucci, and S. D. Martino. Mining Spatio-Temporal Datasets : Relevance, Challenges and Current Research Directions. In J. Ponce and A. Karahoca, editors, *Data Mining and Knowledge Discovery in Real Life Applications*, pages 215–228. InTech Education and Publishing, Vienna, Austria, January 2009. 14, 18

[131] A. Ketterlin and P. Gançarski. Sequence similarity and multi-date image segmentation. In *4th Intl Workshop on the Analysis of Multitemporal Remote Sensing Images*, Leuven, Belgique, July 2007. 25

[132] A. Knobbe, B. Crémilleux, J. Furnkranz, and M. Scholz. From Local Patterns to Global Models : The LeGo Approach to Data Mining. In *From Local Patterns to Global Models (LeGo-09), ECMLPKDD'09 Workshop, Bled, Slovenia, 7 September 2009.* 28, 29, 146

[133] Y. Kodratoff. Techniques et outils de l'extraction de connaissances à partir des données. *Signaux*, 92 :38–43, March 1998. 10

[134] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps.* Springer-Verlag New York, Inc., third edition, 2001. 12

[135] A. Konar. *Computational Intelligence : Principles, Techniques and Applications.* Springer-Verlag, Berlin, Germany, 2005. 27

[136] I. Kopanakis and B. Theodoulidis. Visual data mining modeling techniques for the visualization of mining outcomes. *Journal of Visual Languages and Computing*, 14(6) :543–589, December 2003. Special Issue on Visual Data Mining. 14

[137] K. Koperski, J. Han, and J. Adhikary. Mining knowledge in geographical data. *Communications of the ACM*, 26, 1998. 14, 18

[138] M. Kryszkiewicz and M. Gajek. Concise representation of frequent patterns based on generalized disjunction-free generators. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02)*, volume 2336 of *Lecture Notes in Computer Science*, pages 159–171, Taipei, Taiwan, 2002. Springer-Verlag. 28

[139] H.-C. Kum, J. Pei, W. Wang, and D. Duncan. ApproxMAP : Approximate Mining of Consensus Sequential Patterns. In *Proceedings of the 3rd International Conference on Data Mining*, San Francisco, CA, USA, 2003. 43

[140] C. Largouët and M.-O. Cordier. Improving the landcover classification using domain knowledge. *AI Communications*, 14(1) :35–43, 2001. 24

[141] C. Lauvernet, F. Baret, and F. Le Dimet. Assimilating high temporal frequency SPOT data to describe canopy functioning : the ADAM project. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2003)*, volume 5, pages 3184–3186, Toulouse, France, July 2003. 80

[142] S. Laxman and P. S. Sastry. A survey of temporal data mining. *SADHANA, Academy Proceedings in Engineering Sciences*, 31 :173–198, April 2006. 16

[143] S. Laxman, P. S. Sastry, and K. P. Unnikrishnan. Discovering Frequent Episodes and Learning Hidden Markov Models : A Formal Connection. *IEEE Transactions on Knowledge and Data Engineering*, 17(11), November. 15

[144] F.-X. Le Dimet and J. Blum. Assimilation de données pour les fluides géophysiques. *MATAPLI, Bull. SMAI*, 67 :35–55, January 2002. 80

[145] C. Le Men. *Segmentation spatio-temporelle d'une séquence temporelle d'images satellitaires à haute résolution.* PhD thesis, Télécom ParisTech, École Nationale Supérieure des Télécommunications, Paris, France, Septembre 2009. 22

[146] C. Le Men, A. Julea, N. Méger, M. Datcu, P. Bolon, and H. Maître. Radiometric evolution classification in high resolution satellite image time series (SITS). In *Proc. of the 5th*

Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security (ESA-EUSC 2008), ESRIN - Frascati, Italy, March 2008. 23, 35, 56, 152, 168

[147] C. Le Men, H. Maître, and M. Datcu. Minimum description length applied to the spatio-temporal segmentation of high resolution satellite image time series. *Telecom Technical Report*, May 2008. 168

[148] M. Leleu, N. Méger, and C. Rigotti. Extraction de motifs séquentiels fréquents sous contraintes dans des données contenant des répétitions. *Revue Ingénierie des Systèmes d'Information (ISI)*, 9(1) :133–159, 2004. 39, 41

[149] L. Li and M. Leung. Robust change detection by fusing intensity and texture differences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 2001. 26

[150] M. Lin and S. Lee. Improving the Efficiency of Interactive Sequential Pattern Mining by Incremental Pattern Discovery. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, Big Island, USA, 2003. 42

[151] D. Lu, P. Mausel, E. Brondizio, and E. Moran. Change detection techniques. *Intl. J. of Remote Sensing*, 25(12) :2365–2407, 2004. 26

[152] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy. Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sensing of Environment*, 105(2) :142–154, 2006. 26

[153] H. Maître. *Le traitement des Images de Radar à Synthèse d'Ouverture*. Hermès Science Publications, Paris, 2001. 126

[154] H. Mannila and H. Toivonen. Discovering Generalized Episodes Using Minimal Occurrences. In *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining (KDD'96)*, pages 146–151. ACM Press, 1996. 15

[155] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 189–194, Portland, USA, 1996. AAAI Press. 28

[156] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3) :259–289, 1997. 15, 47

[157] F. Masseglia, F. Cathala, and P. Poncelet. The PSP approach for mining sequential patterns. In *Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery in Databases (PKDD'98)*, volume 1510, pages 176–184, Nantes, France, September 1998. LNAI, Springer Verlag. 27, 39, 40

[158] F. Masseglia, P. Poncelet, and M. Teisseire. Incremental Mining of Sequential Pattern in Large Databases. *Data and Knowledge Engineering*, 46(1) :97–121, 2003. 42

[159] N. Méger. *Recherche automatique des fênetres temporelles optimales des motifs séquentiels*. PhD thesis, L'Institut National des Sciences Appliquées de Lyon, Décembre 2004. 35, 46

[160] N. Méger, R. Jolivet, C. Lasserre, F. Lodge, E. Trouvé, M.-P. Doin, S. Guillaso, A. Julea, P. Bolon, and C. Rigotti. Spatio-Temporal Mining of ENVISAT SAR Interferogram Time Series over the Haiyuan Fault in China. In *Proceeding of the 6th International Workshop on the Analysis of Multi-Temporal Remote Sensing Images*, Trento, Italy, 2011. DOI :10.1007/978-3-642-23184-1_15. 59, 88, 145, 150

[161] H. Miller. Geographic data mining and knowledge discovery. In J. P. Wilson and A. S. Fotheringham, editors, *Handbook of Geographic Information Science*, pages 352–366. Blackwell Publishing Ltd, 2008. 17

[162] H. Miller and J. Han. Geographic data mining and knowledge discovery : an overview. In H. Miller and J. Han, editors, *Geographic data mining and knowledge discovery*, pages 3–32. Taylor & Francis, 2001. 14, 17

[163] H. J. Miller and E. A. Wentz. Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93(3) :574–594, 2003. 17

[164] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2) :203–226, 1982. 44

[165] K. Morik, J.-F. Boulicaut, and A. Siebes, editors. *Local Pattern Detection*, volume 3539 of *Lecture Notes in Computer Science*. Springer, 2005. 28, 29

[166] V. I. Myers. Soil, water and plant relations. In J. R. Shay, editor, *Remote Sensing with special reference to agriculture and forestry*, pages 253–297. National Academy of Sciences, Washington D. C., 1970. 172

[167] M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.*, 27(3) :267–289, November 2006. 32

[168] E. Nezry, G. Genovese, G. Solaas, and S. Rémondière. ERS - Based early estimation of crop areas in Europe during winter 1994-95. In G. T.-D., editor, *ERS Application, Proceedings of the Second International Workshop held 6-8 December 1995 in London*, volume 383 of *ESA Special Publication*, page 13, 1996. 21

[169] R. Ng, L. Lakshmanan, and J. Han. Exploratory Mining and Pruning Optimizations of Constrained Association Rules. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM'98)*, pages 13–24. ACM Press, 1998. 28, 40, 47

[170] E. Oja. Self-organising maps and computer vision. In H. Wechsler, editor, *Neural Networks for Perception (Human and Machine Perception)*, volume 1, pages 368–385. Academic Press, 1992. 24

[171] F. Oro, F. Baret, and R. Vintila. Evaluating of SPOT/HRV data over temporal series acquired during the ADAM project. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2003)*, volume 4, pages 2209–2211, Toulouse, France, July 2003. 80

[172] Y.-H. Pao. *Adaptative Pattern recognition and Neural Networks*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. 24

[173] S. Parthasarathy, M. Zaki, M. Ogihara, and S. Dwarkadas. Incremental and Interactive Sequence Mining. In *Proceedings of the 8th International Conference on Information and Knowledge Management*, Kansas City, USA, 1999. 42

[174] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1) :25–46, January 1999. 28

[175] J. Pei, B. Han, and W. Wang. Mining Sequential Patterns with Constraints in Large Databases. In *Proc. of the 11th International Conference on Information and Knowledge Management (CIKM'02)*, pages 18–25, McLean, VA, USA, November 2002. ACM Press. 39, 42, 44, 45, 47, 48

[176] J. Pei and J. Han. Can we push more constraints into frequent pattern mining ? In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 350–354, New York, NY, USA, 2000. ACM Press. 47

[177] J. Pei and J. Han. Constrained frequent pattern mining : a pattern-growth view. *ACM SIGKDD Explorations*, 4(1) :31–39, 2002. 45

[178] J. Pei, J. Han, B. Mortazavi-Asl, and H. Pinto. PrefixSpan : Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'01)*, pages 215–226, Heidelberg, Germany, 2001. 15, 34, 39, 41, 161

[179] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(10) :1424–1440, 2004. 27, 64, 86, 88

[180] J. Pei, J. Han, and W. Wang. Constraint-based Sequential Pattern Mining : the Pattern-Growth Methods. *J. Intell. Inf. Syst.*, 28(2) :133–160, 2007. 45, 46, 47, 48

[181] C.-S. Perng, H. Wang, S. Ma, and J. L. Hellerstein. Discovery in multi-attribute data with user-defined constraints. *SIGKDD Explorations*, 4(1) :56–64, 2002. 46

[182] I. Petillot. *Cominaison d'informations hétérogènes : intégration d'images RSO pour la surveillance des glaciers alpins*. PhD thesis, Université de Savoie, Décembre 2008. 127, 136

[183] I. Petillot, E. Trouvé, P. Bolon, A. Julea, Y. Yan, M. Gay, and J.-M. Vanpé. Radar-Coding and Geocoding Lookup Tables for the Fusion of GIS and SAR Data in Mountain Areas. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 7(2) :309–313, 2010. 126

[184] I. Petillot, G. Vasile, E. Trouvé, P. Bolon, M. Gay, M. Koehl, and A. Julea. Rectification radar de données géoréférencées : application a l'analyse de données dans les régions de haute montagne. In *Onzième congrès francophone des jeunes chercheurs en vision par ordinateur ORASIS*, Obernai, France, 2007. 126

[185] F. Petitjean, P. Gançarski, and F. Masseglia. Extraction de motifs d'évolution dans les séries temporelles d'images satellites. In *Spatial Analysis and GEOmatics*, November 2010. 22, 26

[186] F. Petitjean, P. Gançarski, F. Masseglia, and G. Forestier. Analysing satellite image time series by means of pattern mining. In C. F. et al., editor, *11th International Conference on Intelligent Data Engineering and Automated Learning*, volume 6283 of *Lecture Notes in Computer Science*, pages 45–52. Springer, 2010. 22

[187] N. Pettorelli, J. O. Vik, A. Mysterud, J.-M. Gaillard, C. J. Tucker, and N. C. Stenseth. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, 20 :503–510, 2005. 20

[188] D. Peuquet and N. Duan. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International Journal of Geographical Information Systems*, 9 :7–24, 1995. 18

[189] R. Platt and L. Rapoza. An evaluation of an object-oriented paradigm for land use/land cover classification. *Professional Geographer*, 60 :87–100, 2008. 22

[190] H. Rahman and G. Dedieu. SMAC : A Simplified Method for the Atmospheric Correction of satellite measurements in the solar spectrum. *International Journal of Remote Sensing*, 16(1) :123–143, 1994. 80

[191] C. Raïssi, P. Poncelet, and M. Teisseire. Speed : Mining maximal sequential patterns over data streams. In *Proceedings of the 3rd International Conference on Intelligent Systems*, pages 546–552, Prague, Czech Republic, 2006. 28

[192] Rigotti, C. DMT4SP : Data Mining Tool 4 Sequential Patterns. online. http ://liris.cnrs.fr/ crigotti/dmt4sp.html. 64

[193] J. F. Roddick and B. G. Lees. Paradigms for spatial and spatio-temporal data mining. In H. Miller and J. Han, editors, *Geographic Data Mining and Knowledge*. Taylor & Francis, 2001. 14, 17, 18

[194] J. F. Roddick and M. Spiliopoulou. A Survey of Temporal Knowledge Discovery Paradigms and Methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4) :750–767, 2002. 14, 15

[195] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring vegetation systems in the Great Plains with ERTS. In *Proceedings of the Third Earth Resources Technology Satellite Symposium*, volume 1, pages 301–317, 1974. 82

[196] L. Schouten, H. van Leeuwen, E. van Valkengoed, J. Desprats, C. King, N. Baghdadi, L. Prévot, N. Bruguier, M. Dechambre, and R. Valentin. Land use classification based on time series of micro-wave data (ERS, Radarsat). study in framework of the project ReSeDa - Assimilation of Multisensor & Multitemporal Remote Sensing Data to Monitor Soil & Vegetation Functioning, 2000. 21

[197] R. Schowengerdt. Soft classification and spatial-spectral mixing. In E. Binaghi, P. Brivio, , and A. Rampini, editors, *Soft Computing in Remote Sensing Data Analysis*, pages 1–6. World Scientific, Singapore, 1996. 21

[198] R. A. Schowengerdt, editor. *Remote Sensing. Models and Methods for Image Processing*. Academic Press, San Diego, CA, USA, second edition, 1997. 2

[199] S. Shekhar, C.-T. Lu, and P. Zhang. A unified approach to detecting spatial outliers. *GeoInformatica*, 7 :139–166, June 2003. 17

[200] M. I. Skolnik, editor. *Radar Handbook*. McGraw-Hill, New York, USA, second edition, 1990. 127

[201] A. Soulet. *Un cadre générique de découverte de motifs sous contraintes fondées sur des primitives*. PhD thesis, Université de Caen / Basse-Normandie, Caen, France, Novembre 2006. 43, 58

[202] A. Soulet and B. Crémilleux. Optimizing constraint-based mining by automatically relaxing constraints. In *The Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 777–780, Houston, USA, 2005. 8

[203] A. Soulet and B. Crémilleux. Mining constraint-based patterns using automatic relaxation. *Intell. Data Anal.*, 13(1) :109–133, 2009. 2, 48

[204] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In H. V. Jagadish and I. S. Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 1–12. ACM Press, 1996. 44

[205] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. In *Proc. of the 5th International Conference on Extending Database Technology (EDBT'96)*, pages 3–17, Avignon, France, September 1996. 34, 39, 40, 46

[206] R. Srikant, Q. Vu, and R. Agrawal. Mining Association Rules with Item Constraints. In *Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining*, pages 67–73, Newport Beach, CA, USA, August 1997. ACM Press. 45

[207] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, May 2005. 109

[208] E. Trouvé, G. Vasile, M. Gay, L. Bombrun, P. Grussenmeyer, T. Landes, J.-M. Nicolas, P. Bolon, I. Petillot, A. Julea, L. Valet, J. Chanussot, and M. Koehl. Combining airborne photographs and spaceborne SAR data to monitor temperate glaciers. Potentials and limits. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4) :905–923, 2007. 127

[209] E. Trouvé, G. Vasile, M. Gay, P. Grussenmeyer, J. Nicolas, T. Landes, M. Koehl, J. Chanussot, and A. Julea. Combining Optical and SAR Data to Monitor Temperate Glaciers. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005)*, volume IV, pages 2637–2640, Seoul, Korea, 2005. 127

[210] I. Tsoukatos and D. Gunopulos. Efficient mining of spatiotemporal patterns. In C. S. Jensen, M. Schneider, B. Seeger, and V. J. Tsotras, editors, *Proceedings of the 7th International Symposium, Advances in Spatial and Temporal Databases (SSTD 2001), Redondo Beach, CA, USA, July 12-15, 2001*, volume 2121 of *Lecture Notes in Computer Science*, pages 425–442. Springer, 2001. 32

[211] C. J. Tucker. Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, 8 :127–150, 1979. 82

[212] C. J. Tucker and P. J. Sellers. Satellite remote sensing of primary production. *International Journal of Remote Sensing*, 7(11) :1395–1416, 1986. 82

[213] G. Vasile, I. Petillot, A. Julea, E. Trouvé, P. Bolon, L. Bombrun, M. Gay, T. Landes, P. Grussenmeyer, and J. Nicolas. High Resolution SAR Interferometry : influence of local topography in the context of glacier monitoring. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2006)*, pages 4008–4011, Denver, CO, USA, 2006. 127

[214] A. Viña, F. R. Echavarria, and D. C. Rundquist. Satellite change detection analysis of deforestation rates and patterns along the Colombia-Ecuador border. *AMBIO : A Journal of the Human Environment*, 33 :118–125, 2004. 25

[215] K. Wang, Y. Jiang, and L. V. S. Lakshmanan. Mining unexpected rules by pushing user dynamics. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*, pages 246–255, New York, NY, USA, 2003. ACM. 46

[216] T. Warner and M. Shank. An evaluation of the potential for fuzzy classification of multispectral data using artificial neural networks. *Photogrammetric Engineering and Remote Sensing*, 63 :1285–1294, 1997. 21

[217] L.-C. Wuu, T.-J. Liu, and K.-M. Chen. A longest prefix first search tree for ip lookup. *Comput. Networks*, 51(12) :3354–3367, 2007. 161

[218] X. Yan, J. Han, and R. Afshar. Clospan : Mining closed sequential patterns in large datasets. In *Proceedings of the SIAM International Conference on DataMinig (SDM'03)*, pages 71–82, San Francisco, CA, USA, May 2003. 28

[219] X. Yao. Research issues in spatio-temporal data mining. In *Workshop on Geospatial Visualization and Knowledge Discovery, Virginia, USA, 18-20 November, 2003.* 18

[220] M. Yuan. Use of knowledge acquisition to build wildfire representation in geographical information systems. *International Journal of Geographical Information Science*, 11 :723–745, December 1997. 18

[221] M. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3) :372–390, 2000. 34

[222] M. Zaki. Sequence Mining in Categorical Domains : Incorporating Constraints. In *Proc. of the 9th International Conference on Information and Knowledge Management (CIKM'00)*, pages 422–429, Washington, DC, USA, November 2000. 39, 41, 44, 46, 159

[223] M. Zaki. SPADE : an efficient algorithm for mining frequent sequences. *Machine Learning Journal, Special issue on Unsupervised Learning*, 42(1/2) :31–60, Jan/Feb 2001. 34, 35, 39, 41, 46, 86, 88

[224] M. J. Zaki. Efficient Enumeration of Frequent Sequences. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*, pages 68–75, New York, NY, USA, 1998. ACM Press. 15, 41

[225] M. J. Zaki. SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, 42(1/2) :31–60, January/February 2001. 27

[226] M. J. Zaki and M. Ogihara. Theoretical foundations of association rules. In *Proceedings of the SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'98)*, pages 1–8, June 1998. 28

[227] K. Zeitouni. Data mining spatial. *Numéro spécial de la Revue internationale de géomatique*, 9(4), 1999. 16

Publications de l'auteur

Revues d'audience internationale

1. A. Julea, N. Méger, C. Rigotti, E. Trouvé, R. Jolivet, Ph. Bolon, "Efficient Spatiotemporal Mining of Satellite Image Time Series for Agricultural Monitoring", Transactions on Machine Learning and Data Mining, Vol. 4, No. 2, October 2011, ibai publishing, Germany.

2. A. Julea, N. Méger, Ph. Bolon, C. Rigotti, M-P. Doin, C. Lasserre, E. Trouvé, V. Lăzărescu, "Unsupervised Spatiotemporal Mining of Satellite Image Time Series Using Grouped Frequent Sequential Patterns", IEEE Transactions in Geoscience and Remote Sensing, Vol. 49, No. 4, 2011, pp. 1417-1430.

3. I. Pétillot, E. Trouvé, Ph. Bolon, A. Julea, Y. Yan, M. Gay, J.-M. Vanpé, "Radar-Coding and Geocoding Lookup Tables for the Fusion of GIS and SAR Data in Mountain Areas", IEEE Geoscience and Remote Sensing Letters (GRSL), Vol. 7, Issue 2, 2010, pp. 309-313.

4. E. Trouvé, G. Vasile, M. Gay, L. Bombrun, P. Grussenmeyer, T. Landes, J.-M. Nicolas, Ph. Bolon, I. Pétillot, A. Julea, L. Valet, J. Chanussot, M. Koehl, "Combining airborne photographs and spaceborne SAR data to monitor temperate glaciers. Potentials and limits", IEEE Transactions in Geoscience and Remote Sensing, Vol. 45, No. 4, 2007, pp. 905-923.

Conférences d'audience internationale avec actes

5. A. Julea, N. Méger, C. Rigotti, E. Trouvé, Ph. Bolon, V. Lăzărescu, "Mining Pixels Evolutions in Satellite Image Time Series for Agricultural Monitoring", in Advances in Data Mining. Applications and Theoretical Aspects - 11th Industrial Conference on Data Mining (ICDM 2011), New York, USA, August 30 - September 3rd, 2011, Lecture Notes in Computer Science, Volume 6870/2011, pages 189-203, DOI :10.1007/978-3-642-23184-1_15.

6. A. Julea, F. Ledo, N. Méger, E. Trouvé, Ph. Bolon, C. Rigotti, R. Fallourd, J.-M. Nicolas, G. Vasile, M. Gay, O. Harant, L. Ferro-Famil, F.Lodge, "POLSAR RADARSAT-2 Satellite Image Time Series Mining Over The Chamonix Mont-Blanc Test Site", IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2011), Vancouver, Canada, July 24-29, 2011, pages 1191-1194.

7. N. Méger, R. Jolivet, C. Lasserre, E. Trouvé, C. Rigotti, F. Lodge, M-P. Doin, S. Guillaso, A. Julea, Ph. Bolon, "Spatiotemporal mining of Envisat SAR interferogram time series over the Haiyuan fault in China", 6th International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, Trento, Italy, July 12-14, 2011, 4 pages, DOI :10.1109/Multi-temp.2011.6005067.10.

8. A. Julea, N. Méger, E.Trouvé, Ph. Bolon, C. Rigotti, R. Fallourd, J.-M. Nicolas, G. Vasile, M. Gay, O. Harant, L. Ferro-Famil, "Spatio-temporal mining of POLSAR satellite image time series", ESA Living Planet Symposium, June 28th - July 2nd, 2010, Bergen, Norway, 6 pages, CD-ROM.

9. A. Julea, N. Méger, C. Rigotti, M-P. Doin, C. Lasserre, E. Trouvé, Ph. Bolon, V. Lăzărescu, "Extraction of frequent grouped sequential patterns from satellite image time series", IEEE International Geoscience And Remote Sensing Symposium (IGARSS 2010), Honolulu, Hawaii, USA, July 2010, pp. 3434-3437.

10. A. Julea, N. Méger, E. Trouvé, Ph. Bolon, "On extracting evolutions from satellite image time series", IEEE International Geoscience And Remote Sensing Symposium (IGARSS 2008), Geospatial Standards & Services II Session, Boston, MA, USA, July 2008, 4 pages (V228 - V231), CD-ROM.

11. C. Le Men, A. Julea, N. Méger, M. Datcu, Ph. Bolon, H. Maître, "Radiometric evolution classification in a High Resolution Satellite Image Time Series (SITS)", ESA-EUSC 2008 - Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security, ESRIN, Frascati, Italy, March 4-6, 2008, 5 pages, CD-ROM.

12. A. Julea, N. Méger, Ph. Bolon, "On mining pixel based evolution classes in satellite image time series", ESA-EUSC 2008 - Conference on Image Information Mining : pursuing automation of geospatial intelligence for environment and security, ESRIN, Frascati, Italy, March 4-6, 2008, 6 pages, CD-ROM.

13. A. Julea, N. Méger, E. Trouvé, "On mining METEOSAT and ERS multitemporal images", ESA-EUSC 2006 - 4th Conference on Image Information Mining for security and Intelligence, Session Theory, November 27-29, 2006, Madrid, Spain, 6 pages, CD-ROM.

14. A. Julea, N. Méger, E. Trouvé, "Sequential patterns extraction in multitemporal satellite images", 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Workshop on Practical Data Mining : Applications, Experiences and Challenges, September 18-22, 2006, Berlin, Germany, CD-ROM.

15. A. Julea, I. Pétillot, G. Vasile, E. Trouvé, V. Buzuloiu, D. Haşegan, "Slant Range Rectification Of Georeferenced Information For SAR Data Analysis In Mountainous Regions", 1st International Summer School "Optoelectronic Techniques for Environmental Monitoring and Risk Assessment", July 31 - August 09, 2006, Baia Mare, Romania, pp. 253-258.

16. A. Julea, G. Vasile, I. Pétillot, E. Trouvé, M. Gay, J.-M. Nicolas, Ph. Bolon, "Simulation of SAR Images and Radar Coding of Georeferenced Information for Temperate Glacier Monitoring", International Conference on Optimization of Electrical and Electronic Equipment, vol. IV, Braşov, Romania, May 2006, pp. 175-180.

17. G. Vasile, I. Pétillot, A. Julea, E. Trouvé, Ph. Bolon, L. Bombrun, M. Gay, T. Landes, P. Grussenmeyer, J.-M. Nicolas, "High Resolution SAR Interferometry : influence of local topography in the context of glacier monitoring", IEEE International Geoscience And Remote Sensing Symposium (IGARSS 2006), Denver, USA, July 2006, 4 pages.

18. E. Trouvé, G. Vasile, M. Gay, P. Grussenmeyer, J.-M. Nicolas, T. Landes, M. Koehl, J. Chanussot, A. Julea, "Combining Optical and SAR Data to Monitor Temperate Glaciers", IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2005), vol. IV, Seoul, Korea, July 2005, pp. 2637-2640.

Conférence d'audience nationale et francophone avec actes

19. I. Pétillot, G. Vasile, E. Trouvé, Ph. Bolon, M. Gay, M. Koehl, A. Julea, "Rectification radar de données géoréférencées : application à l'analyse de données dans les régions de haute montagne", Onzième congrès francophone des jeunes chercheurs en vision par ordinateur, ORASIS, 4 - 8 juin 2007, Obernai, France, 8 pages, CD-ROM.

Publications soumises

20. A. Julea, N. Méger, Ph. Bolon, V. Lăzărescu, "Spatiotemporal mining of evolutions in Satellite Image Time Series", Scientific Bulletin of University Politehnica of Bucharest, Series C : Electrical Engineering and Computer Science, soumis.

Glossaire

ACP analyse en composantes principales. 12

ADAM Assimilation de Données par Agro Modélisation. 80–84, 86–88, 93, 96, 102, 105, 111, 118, 130–132, 141, 142, 149, 150, 155

CG connexité globale. 58, 59, 62, 65, 68, 73, 77, 88, 93, 103, 123, 143–145

CL connexité locale. 73

CM connexité moyenne. 58–60, 62, 65–73, 77, 78, 88, 89, 91, 92, 95–98, 100, 102–104, 123, 132, 133, 143, 144

CRSM connexité relative au support minimum. 3, 60, 62, 65, 66, 68–73, 77, 88, 93–98, 100–104, 123, 132, 143, 144

CS contrainte de support (ou de fréquence). 98, 102, 145

DCT Transformée en Cosinus Discrète 1D (en anglais Discrete Cosine Transform). 152, 153

DFT Transformée de Fourier Discrète (en anglais Discrete Fourier Transform). 152

DMT4SP Data Mining Tool 4 Sequential Patterns. 64

ECD Extraction de Connaissances à partir des Données (en anglais Knowledge Discovery in Database, KDD). 1, 3, 7, 10, 11, 13, 17, 18, 27, 49, 73, 141, 144

EFIDIR Extraction et Fusion d'Informations pour la mesure de Déplacements par Imagerie Radar. 64, 126, 128, 136

EM Espérance-Maximisation (en anglais Expectation Maximization). 22, 56, 152

ENVISAT ENVironmental SATellite. 128

ERS European Remote Sensing satellite. 35, 128, 136, 150, 159

ESA Agence Spatiale Européenne. 128

FDS Fouille de Données Spatiales (en anglais Spatial Data Mining). 14, 16, 17

FDS-T Fouille de Données Spatio-Temporelles (en anglais Spatio-Temporal Data Mining). 14

FDT Fouille de Données Temporelles (en anglais Temporal Data Mining). 14

FreeSpan FREquEnt pattern-projected Sequential PAtterN mining. 41

GSP Generalized Sequential Pattern. 40, 42

INRDA Institut National de Recherche et Développement de l'Agriculture (en roumain Institutul Naţional de Cercetare Dezvoltare Agricolă, INCDA). 81

InSAR Interférométrie Radar à Synthèse d'Ouverture (en anglais Interferometric Synthetic Aperture Radar). 159

ISE Incremental Sequence Extraction. 42

ISM Incremental Sequence Mining. 42

IVA Indice de Végétation Amélioré (en anglais Enhanced Vegetation Index, EVI). 26

IVDN Indice de Végétation Différentielle Normalisée (en anglais Normalized Difference Vegetation Index, NDVI). 4, 12, 26, 82, 83, 87, 91, 109, 111, 112, 122, 149, 165, 167, 173

KIM Knowledge driven Image Information Mining. 30

KISP Knowledge base assisted Incremental Sequential Pattern. 42

MDL Longueur de Description Minimale (en anglais Minimum Description Length). 168

METEOSAT satellites météorologiques géostationnaires réalisés sous maîtrise d'oeuvre de l'Agence Spatiale Européenne (ESA). 35, 150, 159

MODIS Moderate Resolution Imaging Spectroradiometer. 26

MS motifs séquentiels. 38, 83, 84, 88, 104

MSF motifs séquentiels fréquents. 3, 7, 8, 21, 34, 35, 37, 38, 49, 62, 65, 66, 69, 70, 72, 73, 77, 83, 86, 88, 91–94, 104, 131, 132, 141, 144, 155, 159

MSFG motifs séquentiels fréquents groupés. 2–4, 65–73, 77, 78, 83, 88–92, 98, 99, 104, 106, 109, 111, 112, 115, 116, 119, 126, 129, 131–133, 135, 138–142, 144, 145

PG Pattern Growth. 42, 47, 64

PIR proche infrarouge. 82, 83, 91, 111, 112, 165, 171, 173

PolSAR Radar à Synthèse d'Ouverture Polarimétrique (en anglais Polarimetric Synthetic Aperture Radar). 136–138, 140, 152

PrefixSpan PREFIX projected Sequential PAtterN mining. 41, 42

PSP Prefix tree for Sequential Pattern. 40

RADAR RAdio Detection And Ranging. 126

RSO Radar à Synthèse d'Ouverture (en anglais Synthetic Aperture Radar, SAR). 21, 126–129, 135, 136, 150

SFG séquentiels fréquents groupés. 113, 116, 139

SOM cartes adaptatives de Kohonen (en anglais Self Organizing Maps). 12, 22, 23

SOTAG graphe d'adjacence temporelle des objets spatiaux (en anglais Spatial Object Temporal Adjancency Graph). 23

SPADE Sequential PAttern Discovery using Equivalence classes. 35, 41, 42, 46

SPAM Sequential PAttern Mining. 41

SPATPAM SPAtio-TemPorAl Mining. 64

SPIRIT Sequential Pattern mining with Regular Expressions. 40, 46

SPOT Satellites Pour l'Observation de la Terre. 35, 80, 82, 83, 111

STIS Série Temporelle d'Images Satellitaires. 1–4, 7, 8, 15, 20–24, 26–28, 30–32, 34, 35, 49, 53, 56, 64, 67, 72, 74, 77, 78, 80–84, 86, 88, 93, 104, 111, 113, 115, 118, 126–128, 130–133, 135, 136, 141–145, 149, 150, 152, 155, 157, 159, 162, 167, 168